A Companion to Analytic Philosophy

Blackwell Companions to Philosophy

This outstanding student reference series offers a comprehensive and authoritative survey of philosophy as a whole. Written by today's leading philosophers, each volume provides lucid and engaging coverage of the key figures, terms, topics, and problems of the field. Taken together, the volumes provide the ideal basis for course use, representing an unparalleled work of reference for students and specialists alike.

Already published in the series

1. The Blackwell Companion to Philosophy
   Edited by Nicholas Bunnin and Eric Tsui-James
2. A Companion to Ethics
   Edited by Peter Singer
3. A Companion to Aesthetics
   Edited by David Cooper
4. A Companion to Epistemology
   Edited by Jonathan Dancy and Ernest Sosa
5. A Companion to Contemporary Political Philosophy
   Edited by Robert E. Goodin and Philip Pettit
6. A Companion to Philosophy of Mind
   Edited by Samuel Guttenplan
7. A Companion to Metaphysics
   Edited by Jaegwon Kim and Ernest Sosa
8. A Companion to Philosophy of Law and Legal Theory
   Edited by Dennis Patterson
9. A Companion to Philosophy of Religion
   Edited by Philip L. Quinn and Charles Taliaferro
10. A Companion to the Philosophy of Language
    Edited by Bob Hale and Crispin Wright
11. A Companion to World Philosophies
    Edited by Eliot Deutsch and Ron Bontekoe
12. A Companion to Continental Philosophy
    Edited by Simon Critchley and William Schroeder
13. A Companion to Feminist Philosophy
    Edited by Alison M. Jaggar and Iris Marion Young
14. A Companion to Cognitive Science
    Edited by William Bechtel and George Graham
15. A Companion to Bioethics
    Edited by Helga Kuhse and Peter Singer
16. A Companion to the Philosophers
    Edited by Robert L. Arrington
17. A Companion to Business Ethics
    Edited by Robert E. Frederick
18. A Companion to the Philosophy of Science
    Edited by W. H. Newton-Smith
19. A Companion to Environmental Philosophy
    Edited by Dale Jamieson
20. A Companion to Analytic Philosophy
    Edited by A. P. Martinich and David Sosa

Forthcoming

A Companion to Genethics
Edited by John Harris and Justine Burley

A Companion to African-American Philosophy
Edited by Tommy Lott and John Pittman

A Companion to African Philosophy
Edited by Kwasi Wiredu

A Companion to Ancient Philosophy
Edited by Mary Louise Gill

A Companion to Early Modern Philosophy
Edited by Steven Nadler

A Companion to Philosophical Logic
Edited by Dale Jacquette

A Companion to Medieval Philosophy
Edited by Jorge J. E. Gracia, Greg Reichberg, and Timothy Noone

Blackwell
Companions to
Philosophy

A Companion to Analytic Philosophy

Edited by

A. P. MARTINICH
and
DAVID SOSA

BLACKWELL
Publishers

First published 2001

2 4 6 8 10 9 7 5 3 1

Blackwell Publishers Inc.
350 Main Street
Malden, Massachusetts 02148
USA

Blackwell Publishers Ltd
108 Cowley Road
Oxford OX4 1JF
UK

Library of Congress Cataloging-in-Publication Data

A companion to analytic philosophy / edited by A. P. Martinich and David Sosa.
p. cm. – (Blackwell companions to philosophy)
Includes bibliographical references and index.
ISBN 0-631-21415-1 (hb : alk. paper)
1. Analysis (Philosophy) 2. Philosophy, Modern – 19th century. 3. Philosophy, Modern – 20th century. I. Martinich, Aloysius. II. Sosa, David, 1966– III. Series.
B808.5 .C555 2001
146′.4 – dc21

00-050770

British Library Cataloguing in Publication Data
A CIP catalogue record for this book is available from the British Library.

Typeset in 10 on 12.5 pt Photina
by Best-set Typesetter Ltd., Hong Kong
Printed in Great Britain by MPG Books Ltd, Bodmin, Cornwall

This book is printed on acid-free paper.

Contents

Contributors

C. Anthony Anderson
Professor of Philosophy, University of California, Santa Barbara

Brad Armendt
Associate Professor of Philosophy, Arizona State University

Thomas Baldwin
Professor of Philosophy, University of York, England

Max Cresswell
Professor of Philosophy, Victoria University of Wellington, New Zealand

Norman Daniels
Professor of Philosophy and Goldthwaite Professor of Rhetoric, Tufts University

James Dreier
Associate Professor of Philosophy, Brown University

Sir Michael Dummett
Wykeham Professor of Logic Emeritus, University of Oxford

Richard Foley
Professor of Philosophy and Dean of the Faculty of Arts and Sciences, New York University

Carl Ginet
Professor Emeritus of Philosophy, Cornell University

Richard Grandy
Carolyn and Fred McManis Professor of Philosophy, Rice University

P. M. S. Hacker
Fellow of St. John's College, University of Oxford

John Heil
Professor of Philosophy, Davidson College

Peter Hylton
Professor of Philosophy, University of Illinois, Chicago

Frank Jackson
Professor of Philosophy in the Philosophy Program, Research School of Social Sciences, and Director of the Institute of Advanced Studies, Australian National University

Philip Kitcher
Professor of Philosophy, Columbia University

Gavin Lawrence
Associate Professor of Philosophy, University of California, Los Angeles

Ernest LePore
Professor of Philosophy and Director of the Center for Cognitive Science, Rutgers University

Peter Ludlow
Professor of Philosophy, State University of New York, Stony Brook

A. P. Martinich
Roy Allison Vaughan Centennial Professor of Philosophy, Professor of History, University of Texas, Austin

Alexander Miller
Senior Research Fellow, Cardiff University

Anselm Winfried Müller
Professor of Philosophy, University of Trier, Germany

Stephen Neale
Professor of Philosophy, Rutgers University

W. H. Newton-Smith
Fairfax Fellow, Jowett Fellow, Jowett Lecturer and Tutor in Philosophy, Balliol College, University of Oxford

Georges Rey
Professor of Philosophy, University of Maryland, College Park

Jay F. Rosenberg
Taylor Grandy Professor of Philosophy, University of North Carolina, Chapel Hill

Sahotra Sarkar
Associate Professor of Philosophy, University of Texas, Austin

Israel Scheffler
Professor Emeritus of Philosophy and Professor Emeritus and Director, Philosophy of Education Research Center, Graduate School of Education, Harvard University

John R. Searle
Mills Professor of Philosophy, University of California, Berkeley

Scott Shapiro
Assistant Professor of Law, Benjamin Cardozo School of Law, Yeshiva University

Walter Sinnott-Armstrong
Professor of Philosophy, Dartmouth College

P. F. Snowdon
Lecturer in Philosophy and Fellow, Exeter College, University of Oxford

David Sosa
Associate Professor of Philosophy, University of Texas, Austin

Ernest Sosa
Romeo Elton Professor of Natural Theology, Brown University, and Professor of Philosophy, Rutgers University

T. L. S. Sprigge
Honorary Fellow, University of Edinburgh

Robert Stalnaker
Professor of Philosophy, Massachusetts Institute of Technology

Frederick Stoutland
Professor Emeritus of Philosophy, St. Olaf College, and Permanent Visiting Professor, University of Uppsala

Avrum Stroll
Research Professor, University of California, San Diego

James van Cleve
Professor of Philosophy, Brown University

Michael Williams
Professor of Philosophy, Johns Hopkins University

Dean Zimmerman
Associate Professor of Philosophy, Syracuse University

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

Introduction

A. P. MARTINICH

Though analytic philosophy was practiced by Plato and reinvigorated in the modern era by René Descartes and Thomas Hobbes among others, we are concerned with it only in its twentieth-century forms. As such, it was revived in two centers, Germany and England. In Germany, Gottlob Frege was exploring the foundations of mathematics and logic. His efforts introduced new standards of rigor that made their way into analytic philosophy generally, through the work of Bertrand Russell and Ludwig Wittgenstein. His discussions of the nature of language and reasoning have also become powerful tools in the hands of later philosophers. Among Frege's many books and articles, the *Grundgesetze*, *Begriffsschrift*, "On Sense and Reference" ("Über Sinn und Bedeutung," 1892) and "Thoughts" ("Gedanken," 1918) stand out as especially significant.

During about the same period in England, G. E. Moore led the way in opposing the then-dominant philosophy of British idealism. While "The Nature of Judgment" is an early criticism of a point in F. H. Bradley's *Logic*, the *locus classicus* of British analytic philosophy is likely "The Refutation of Idealism" (1903), a criticism of the formula *esse est percipi* ("to be is to be perceived"). A crucial part of that argument is Moore's claim that the concept of the sensation of yellow contains two parts: the sensation that is unique to each person and the yellowness that can be perceived by many people. Even when idealists conceded that there was some kind of duality here, they insisted on a kind of inseparability.

To use a general name for the kind of analytic philosophy practiced during the first half of the twentieth century, initially in Great Britain and German-speaking countries, and later in North America, Australia, and New Zealand, "conceptual analysis" aims at breaking down complex concepts into their simpler components. Successive analyses performed on complex concepts would yield simpler concepts. According to Moore, the process might lead ultimately to simple concepts, of which no further analysis could be given. The designation "conceptual" was supposed to distinguish the philosophical activity from various analyses applied to nonconceptual objects. Physics was famous in the twentieth century for breaking down atoms into protons, neutrons, and electrons, and these subatomic particles into an array of more exotic components. And analytic chemistry aims at determining chemical compositions. The analogy between philosophy and science inspired the name "logical atomism," a theory that flourished between

1920 and 1930. Both Wittgenstein and Bertrand Russell maintained that there must be simple, unanalyzable objects at the fundamental level of reality. Wittgenstein thought that the simples existed independently of human experience, Russell that they existed only for as long as one's attention was fixed on them.

Notwithstanding the analogy between scientific and philosophical analysis, most philosophers in the first half of the twentieth century maintained that philosophy was very different from science. In his *Tractatus Logico-Philosophicus* (1921), Wittgenstein wrote: "Philosophy is not one of the natural sciences. (The word 'philosophy' must mean something whose place is above or below the natural sciences, not beside them.)" (4.111). This conveniently left open which was superior.

But if there is anything constant in analytic philosophy, it is change, and the opposite view of the relation between science and philosophy has dominated the second half of the century. Largely owing to the influence of W. V. Quine, many philosophers have come to believe that philosophy is continuous with science. Yesterday's heresy is today's orthodoxy. Whichever view is correct, the division between the philosophical analysis of concepts and the nonphilosophical scientific analysis of nonconceptual objects should perhaps not be taken too strictly. Concepts and hence philosophy would be of no use if they did not make contact with the nonconceptual world. In addition, science uses concepts, many of which may be among the most fundamental of reality. To paraphrase Kant, perceptions without concepts are blind; concepts without perceptions are empty.

Overlapping with the latter period of logical atomism is logical positivism, which may be dated from Moritz Schlick's founding of the Vienna Circle in 1924. One of its principal doctrines was that science is a unity; and one of its principal projects was to show how to translate all meaningful language into scientific language, in other words, to reduce meaningful nonscientific language to scientific language. This project cannot be successful unless something distinguishes meaningful from nonmeaningful expressions. A. J. Ayer probably devoted more energy and displayed more ingenuity in trying to formulate a criterion of meaningfulness than anyone else. His first effort was presented in *Language, Truth and Logic* (1936), the book that became the most widely known statement of logical positivism and which introduced that philosophy to the anglophone public. The basic idea is that a sentence is meaningful if and only if it is either analytic (or contradictory) or empirically verifiable. Various objections were raised to this, and to every revision of this criterion. Part of the problem was the status of the criterion itself. Either it would be analytic and hence vacuous, or it would be empirical but then not completely confirmed. Logical positivism had been dead for some time when it was buried by Carl G. Hempel's "Problems and Changes in the Empiricist Criterion of Meaning" (1950) and W. V. Quine's "Two Dogmas of Empiricism" (1951). Nevertheless, Ayer and others never abandoned the spirit of verifiability.

What had already begun to take the place of logical positivism in the 1940s was ordinary-language philosophy, one strand of which emanated from Cambridge in the later philosophy of Wittgenstein, the other from Oxford. One of Wittgenstein's motivating beliefs was that philosophy creates its own problems, and that means that they are not genuine problems at all. The confusion arises from philosophers' misuse of ordinary words. They take words out of their ordinary context, the only context in which they have meaning, use them philosophically, and thereby discover anomalies with the

displaced concepts expressed by these words: "For philosophical problems arise when language *goes on holiday*." Wittgenstein questioned many of the assumptions of analytic philosophy – from the nature and necessity of analysis to the nature of language – in a discursive and dialectical style so inimitable that it was as if Ludwig were talking to Wittgenstein. His oracular aphorisms, such as "Don't ask for the meaning, ask for the use" and "To understand a sentence is to understand a language" stimulated a variety of reactions, from the Fregean interpretations of Peter Geach and Michael Dummett, to the holism of Quine and Donald Davidson, to the deconstructivist approaches of O. K. Bouwsma and D. Z. Phillips.

The other strand of ordinary-language philosophy came from Oxford, under the leadership of Gilbert Ryle and J. L. Austin. These philosophers, more numerous than the Cambridge group (Antony Flew, J. O. Urmson, and G. J. Warnock, deserve to be mentioned), did not so much think that there were no philosophical problems as say that philosophical problems could be solved through the careful analysis of the distinctions inherent in ordinary language. The purpose of Austin's "Ifs and Cans" and "A Plea for Excuses" was to elucidate the problem of freedom and determinism, which arose from his understanding of Aristotle (see his *Philosophical Papers*, 2nd edition, p. 180). He said that while ordinary language was not the last word in philosophy, it was the first. He certainly was not opposed to philosophers developing theories.

Austin, who had been a closet logical positivist according to A. J. Ayer, coined the term "performative utterance" as part of his refutation of the central thesis of logical positivism, namely, that all sentences that were cognitively meaningful were either true or false. Austin pointed out that some straightforwardly meaningful sentences, sentences that did not contain suspicious words like "beautiful," "good," or "God," were not the kind of sentences that could have a truth-value: "I bequeath my watch to my brother," "I christen this ship the Queen Elizabeth II," and "I bet ten dollars that Cleveland wins the pennant." Although the concept of performatives did the work it was designed to do, the distinction between performatives and "constatives" (roughly, statements) could not be sustained; and Austin replaced that distinction with another, between locutionary, illocutionary, and perlocutionary acts. In the 1960s, John Searle, who was trained at Oxford by ordinary-language philosophers, showed that Austin's latter theory was itself inadequate and replaced it with his own fully-developed theory in *Speech Acts* (1969) and *Expression and Meaning* (1979).

By the late 1960s ordinary language had lost its dominance. Some of the Oxford philosophers were instrumental in its demise. Searle, as mentioned, developed a full-fledged theory of speech acts, and then used it as inspiration for foundational work on the nature of intentionality and the social world. One of his teachers and a colleague of Austin's, H. P. Grice, developed his own theory of language use, a theory complementary in many ways to Searle's.

A more dramatic cause of the demise of ordinary language philosophy is attributable to one of its chief practitioners, P. F. Strawson. In *Individuals* (1959), he resurrected metaphysics, an area of philosophy that was considered unacceptable by logical positivism. Strawson distinguished between "stipulative" (bad) metaphysics and "descriptive" (good) metaphysics. His descriptive project, to lay "bare the most general features of our conceptual structure," was supposed to differ from logical or conceptual analysis only "in scope and generality." At almost the same time, the American W. V. Quine

published *Word and Object* (1960). His approach differed from Strawson's primarily in emphasizing the genesis of the most general concepts and in accommodating itself explicitly to empirical psychology and physics.

Once metaphysics had been made respectable again, philosophers felt more comfortable pursuing a large variety of problems in a variety of ways. Metaphysical systems became more elaborate when Saul Kripke used possible worlds to prove theorems about modal logic. Some subsequent positions can even be thought outlandish, such as David Lewis's view that every possible world exists, and exists in the same sense our own world does – outlandish but not disreputable. Some disciplines that had been relatively neglected between 1930 and 1960 were reinvigorated, for example, ethics and political philosophy by John Rawls, most notably in *A Theory of Justice* (1971); and some questions, such as the meaning of life, were mulled over by, for example, Thomas Nagel in an analytically respectable way. Perhaps two of the most salient characteristics of the period from 1970 onwards were first, the interest of analytic philosophers in the foundations of empirical sciences, from physics through biology to psychology, and second, their use of and contribution to artificial intelligence and cognitive science. Analysis was largely abandoned and replaced by a desire for philosophical doctrines that were variously more intelligible or intellectually respectable to physicists, logicians, or psychologists. This would explain the large presence of philosophers in cognitive science, linguistics, logic, and the philosophy of science; but has perhaps also led to what Searle has called "the rediscovery of the mind" in a book by that name.

There were other consequences of the revival of metaphysics. Some philosophers, respected for their work as early as the 1950s, for example Roderick Chisholm and Wilfrid Sellars, but not closely associated with any of the schools we have mentioned, grew in significance. Some philosophers turned to the history of modern philosophy, notably, Strawson and Jonathan Bennett on Kant, Bennett on Locke, Berkeley, and Hume, and Bernard Williams and Margaret Wilson on Descartes. Some philosophers who became important in the last quarter of the twentieth century, notably Richard Rorty, declared analytic philosophy misconceived, bankrupt, or similarly deficient. In making their position clear and in aiming at cogency, they are analytic philosophers in spite of themselves.

It is likely less helpful to talk about one or another movement in philosophy after 1965. No one method or doctrine dominated. Sometimes a philosopher championing a view became its most significant critic or at least moved on to something quite different, paradigmatically Hilary Putnam. What can be said about the last quarter of the twentieth century is that the original conception of analysis and most of its presuppositions were abandoned by almost all analytic philosophers. Gone is the assumption that concepts of philosophical importance are often composed of simpler sharply-defined concepts. Quine's arguments that there is no principled distinction between analytic and synthetic statements is just a special case of the broader thesis that language and hence thought are essentially indeterminate.

We have been explaining and illustrating analytic philosophy in the last century without defining it. It probably defies definition since it is not a set of doctrines and not restricted in its subject matter. It is more like a method, a way of dealing with a problem, but in fact not one method but many that bear a family resemblance to each other.

Once when Gilbert Harman was asked, "What is analytic philosophy?," he said (tongue firmly in cheek), "Analytic philosophy is who you have lunch with." In general, analytic philosophy has become highly pluralistic and in many ways hardly resembles what was done in the first half of the century. The refectory of analytic philosophy is not as clubby as it once was. Many more people sit at the table, and many more different kinds of food, prepared in more ways, are served. Perhaps what makes current analytic philosophers analytic philosophers is a counterfactual: they would have done philosophy the way Moore, Russell, and Wittgenstein did it if they had been doing philosophy when Moore, Russell, and Wittgenstein were. The multiplicity of analytical styles is one reason for organizing the volume by individual philosopher and not by theme.

Over forty of the greatest analytic philosophers of the last century are discussed in this volume. At least thirty of them, we believe, would be on virtually any sensible list of forty outstanding analytic philosophers. Many other philosophers have almost as good a claim to be included in this volume. To name only some of those who are not alive, the following were considered and finally, reluctantly, not included: Max Black, Gustav Bergmann, Herbert Feigl, Paul Feyerabend, Gareth Evans, C. I. Lewis, J. L. Mackie, Ernest Nagel, H. H. Price, H. A. Prichard, A. N. Prior, Hans Reichenbach, Moritz Schlick, Gregory Vlastos, Friedrich Waismann, and John Wisdom.

Some philosophers were excluded because they do not fit squarely within the tradition of analytic philosophy as ordinarily understood: John Dewey, William James, Charles Sanders Pierce, John Cook Wilson, and, ironically, Alfred North Whitehead, co-author with Russell of one of the century's greatest works of logic, *Principia Mathematica*.

While the reputations of some of the philosophers included are as high as they ever were, e.g. Frege and Russell, those of others have declined, not always justifiably, for example, those of C. D. Broad and Rudolf Carnap. In making our decisions we have tried not to be prejudiced either for or against any school, method, or time period, but to reflect the relative importance of various philosophers over the entire twentieth century.

We know that our selection will be controversial, even though it was influenced by the judgments of many colleagues. A referee of our proposal wrote that the editors seem to "aim at enraging the reader." Most analytic philosophers will believe that some other list of people would have been better. We are sympathetic. Neither of us completely agrees with the final selection. Each believes that at least three other philosophers have a better claim to be included than some that were. In order to preserve "plausible deniability," we have agreed not to comment further on the lists in any written form, and not to appear together at any public gathering of philosophers for five years.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

1

Gottlob Frege (1848–1925)

MICHAEL DUMMETT

Friedrich Ludwig Gottlob Frege was born in Wismar in 1848, and died in Bad Kleinen in 1925. His whole adult career was spent, from 1874 to 1918, in the Mathematics Department of Jena University. He devoted almost all his life to work on the borderline between mathematics and philosophy. In his lifetime, that work was little regarded, save by Bertrand Russell and Ludwig Wittgenstein; his death was marked by very few. Yet today he is celebrated as the founder of modern mathematical logic and as the grandfather of analytical philosophy.

Frege's intellectual career was unusual. Most philosophers and mathematicians make contributions to diverse topics within their fields; but Frege set himself to achieve one particular, though extensive, task. From 1879 to 1906 he pursued this single ambitious aim: to set arithmetic upon secure foundations. Virtually all that he wrote during those years was devoted to this project, or to the elaboration of ideas that he developed in the course of trying to carry it through and took as integral to it. The term "arithmetic," as he used it, is to be understood in a broad sense: for him, it comprised, not only number theory (i.e. the theory of the natural numbers), but also analysis, that is, the theory of real and complex numbers. He did not attempt to construct the foundations of geometry. He viewed mathematics in the traditional way, as divided into the theory of quantity, and thus of cardinal numbers and of numbers measuring the magnitudes of quantities (real numbers), and the theory of space (points, lines, and planes). He believed these two parts of mathematics to rest on different foundations. The foundations of arithmetic – of number theory and analysis – are purely logical. But although the truths of geometry are a priori, they rest upon spatial intuition: they are synthetic a priori, in the Kantian trichotomy Frege accepted. Kant was therefore right about geometry; but he was wrong about arithmetic. All appeal to spatial or temporal intuition must be expelled from arithmetic: its concepts must be formulated and its basic principles established without recourse to intuition of any kind.

It was to the task of establishing the purely logical foundations of arithmetic that Frege devoted his whole intellectual endeavor. In carrying out this task, he was led into some purely philosophical investigations; it is for this reason that, although a mathematician, he is now held in such high regard by philosophers of the analytic school. His attempt to provide arithmetic with secure logical foundations was embodied in three books, all of high importance, although he also published a number of articles,

spin-offs from his central endeavor. The first of these three books was the short *Begriffsschrift* (*Conceptual Notation*) of 1879: this expounded a new formalization of logic. It was the first work of modern mathematical logic: it contained an axiomatic system of predicate logic of precisely the type that was to become standard. The notation was utterly different from that which would become standard, and was essentially two-dimensional, the two clauses of a conditional being written on different lines; but the notation was essentially isomorphic to that which Peano, Hilbert, and Russell later made standard. An English translation of the book and of some related articles was published by T. W. Bynum in 1972. The *Begriffsschrift* received six reviews, including a lengthy one by Ernst Schröder; but none of the reviewers understood Frege's intention or his achievement. The second book was also short, though not quite so short as *Begriffsschrift*. It was called *Die Grundlagen der Arithmetik* (*The Foundations of Arithmetic*) and appeared in 1884. In order that it should be accessible to as many as possible, the book was written without the use of logical symbols. Frege first surveys a range of rival theories on the status of arithmetical propositions and the nature of (cardinal) number, and demolishes them with trenchant arguments; this part of the book is largely philosophical in character. Then, having to his satisfaction left no space for any other theory but his own, he proceeds to sketch a purely logical derivation of the fundamental laws of arithmetic. It is to be doubted whether any other philosophical treatise of comparable length has, since Plato, ever manifested such brilliance as Frege's *Grundlagen*. An English translation by J. L. Austin was published in 1950, and a critical edition by C. Thiel in 1986. This time Frege's book received five reviews, again none of them adequate to their subject matter. One was by Georg Cantor, who unhappily does not seem to have tried to understand the work, with which he might have been expected to be in large sympathy.

Frege had thought that he was on the verge of success in constructing definitive foundations for arithmetic. He had thought that his *Grundlagen* would make this plain to the world of philosophers and mathematicians. He became intensely depressed by his failure to have conveyed to that world the magnitude of his achievement. At the same time, he became aware of deficiencies in the philosophical basis on which, in *Grundlagen*, he had rested his arguments and which underpinned his formal logic. There followed five years during which he published nothing, but engaged in a thoroughgoing revision of his philosophy of logic and of his formal logic. The outcome of this revision he expounded in a lecture, *Function und Begriff* (*Function and Concept*), given in 1891. He then set about a complete formal exposition of his foundations of arithmetic in his third great book, *Grundgesetze der Arithmetik* (*Basic Laws of Arithmetic*). This was utterly unlike *Grundlagen*. The first volume came out in 1893, the second volume not until ten years later, in 1903. The book is incomplete; a third volume must have been planned, but it never appeared. In the first volume of *Grundgesetze*, there is no argument, only exposition. Frege began by explaining his formal logical system, and expounding, without giving any argument for or justification of them, the philosophical, or, more exactly, semantic notions that underpinned it. He provides what is in effect a precise semantic theory for the formal language used in the book. This makes up Part I, of which an English translation by M. Furth was published in 1964.

There follows in Part II a string of formal derivations, carried out in Frege's far from easily read symbolism, which execute in detail the program sketched in *Grundlagen* for

constructing a logical foundation for the theory of cardinal numbers, including the natural numbers, understood as finite cardinals. Part II was concluded in the second volume of 1903, and is followed by a Part III devoted to the real numbers, a topic Frege had never treated in any detail before. Part III is not completed in the second volume; a clear intention to complete it proves that a third volume was contemplated. The first half of Part III is in prose, not symbols; Frege has changed his approach. He attempts in this half to do for real numbers what he had done in *Grundlagen* for cardinal numbers: to review and criticize rival theories so that there would appear no alternative to his own construction of a foundation for analysis, carried out by proof in his formal system in the second half of Part III. Unhappily, he had lost his touch. Whereas in *Grundlagen* nothing is mentioned save what will carry the argument forward, the first half of Part III of *Grundgesetze* reads as if he was merely determined to get his own back on the other theorists who had neglected him, Cantor and Dedekind included. Criticisms are made of features of their work quite irrelevant to the main strand of the argument; errors are pointed out which could easily be rectified, without any indication of how to rectify them. A powerful critique of formalism is included. It is possible to extract Frege's reasons for the strategy he adopts for constructing foundations of analysis; but the whole lacks the brilliance and the exquisite planning of *Grundlagen*.

When he had finished composing the second volume of *Grundgesetze*, Frege must have felt a deep contentment. Still embittered by the neglect of his work, he surely believed that he had attained his life's goal: he had constructed what he thought to be definitive foundations for the theories both of natural numbers and of real numbers. But, while the book was in press, he received the heaviest blow of all, delivered by one of his few admirers, the young Bertrand Russell. Russell wrote in June 1902 to explain the celebrated contradiction he had discovered in the (naive) theory of classes, and to point out that it could be derived in Frege's logical system (see RUSSELL). At first Frege was shattered. Then, as he reflected on the matter, he devised a weakening of his Basic Law V, which governed the abstraction operator used for forming symbols for classes and was responsible for the contradiction; he was confident that this modification would restore consistency to his system. The modification was explained in an Appendix added to Volume II of *Grundgesetze*. The fact was, however, that the modified Basic Law V still allowed the derivation of a contradiction. Whether Frege ever discovered this is uncertain; but what he must have discovered was that, in its presence, none of the proofs he had given of crucial theorems would go through. His wife died in 1905. It took him until August 1906 to convince himself that his logical system could not be patched up; at that point he had to face the fact that the project to which he had devoted his life had failed. *Grundgesetze* remains the only part of Frege's published work of which no full English translation has yet appeared (save for the translation by Furth of Part I, which also contains the Appendix to Volume II, and excerpts in the volume of translations by Geach and Black).

A brief fragment among Frege's literary remains, dated August 1906, asks the question, "What can I regard as the outcome of my work?" In other words, "What remains now that the contradiction has destroyed my logical foundations for arithmetic?" Frege's answer was that what survived of enduring value was his logical system, stripped of the abstraction operator, and the whole structure of philosophical logic

which, from 1891 onwards, had underpinned it. He continued to think about these topics, although it was not for years that he published anything more. Eventually, during the war years, he published the first three chapters of what was planned as a comprehensive treatise on logic, although it was never finished. In the very last years of his life, he finally turned again to the foundations of mathematics; reversing his lifelong view, he began a derivation of arithmetic from geometry, but did not carry it very far.

Frege's judgment of 1906 about where the value of his work lay was at first the judgment of those who participated in the revived study of Frege's writings. In the Preface to his *Tractatus Logico-Philosophicus* of 1922, Ludwig Wittgenstein had written of "the great works of Frege and the works of my friend Bertrand Russell," and had referred frequently to Frege in the book. Despite the celebrity of the *Tractatus* and its wide influence, this failed to stimulate more than a very few philosophers to find out anything about Frege. Rudolf Carnap, who had actually heard Frege's lectures, Alonzo Church, and Peter Geach continued to hold him in high regard (see CARNAP); but the majority of philosophers, British, German, and American, went on ignoring him. The revival of interest in him began in a slow way in the 1960s, and gathered momentum in the 1970s; by the end of that decade, he had become required reading for any student of analytic philosophy. But the interest in his philosophy of arithmetic was meager; it was taken for granted that his system's having run foul of Russell's contradiction destroyed all its pretensions to serious consideration. Interest in Frege therefore concentrated on what made him the grandfather of analytic philosophy: his philosophical analysis of language and of thought which underlay his formal logic.

The key to a modern system of predicate logic is of course the quantifier-variable notation for generality, which Frege introduced for the first time in *Begriffsschrift*. He employed only negation, the conditional, and the universal quantifier; he did not use symbols for conjunction, disjunction, or the existential quantifier, but expressed these by means of the three logical constants for which he did have symbols. Frege insisted that his symbolism, unlike that in the Boolean tradition such as Schröder's, could incorporate a formal language: it needed only the addition of suitable nonlogical constants, predicates, etc., to be capable of framing sentences on any subject matter whatever, and of carrying out deductive reasoning concerning it. Frege did not conceive of formulae in his symbolism in the way that Tarski was to do, namely as built up from atomic formulae containing free variables waiting to become bound in the process of forming complex formulae. Rather, he thought of them as built up out of atomic *sentences*. This required that, before attaching a quantifier, there must first be formed, from a suitable sentence, what we should call a predicate, but, for Frege, was a functional expression or expression for a concept. (He eschewed the term "predicate" as too closely associated with the traditional subject–predicate logic.) Such an expression was "incomplete" or "unsaturated": it could not stand on its own, but had gaps in it, being formed from a sentence by omitting one or more occurrences of a singular term. When a quantifier was attached to it, the bound variables governed by it were to be inserted into these gaps, thereby showing to what expression for a concept the quantifier had been attached.

When he wanted to speak of a particular expression for a concept, Frege used the lower-case Greek letter ξ to indicate the gaps in it; but an expression containing ξ was

no part of the formal language, but only something to be used metalinguistically to speak about the formal language. Thus from the sentence "Pitt respects Pitt's father" (which represents the way we are meant to understand the colloquial "Pitt respects his father") three different expressions for concepts could be formed: "ξ respects Pitt's father," "Pitt respects ξ's father," and "ξ respects ξ's father." This exemplifies the fact that a sentence can be analyzed in different ways. Frege insists in *Begriffsschrift* that these different analyses have nothing to do with the content of the sentence, but only with our way of looking at it; in other words, our grasp of the content of the sentence does not depend upon our noticing that it is possible to analyze it in one way or another, for example that the proper name "Pitt" occurs twice within it. In one sense, each of the three expressions for concepts occurs within the sentence; but none of the different concepts is part of the content of the sentence. By attaching the universal quantifier to these three expressions for concepts, we obtain respectively "For all *a*, *a* respects Pitt's father," "For all *a*, Pitt respects *a*'s father," and "For all *a*, *a* respects *a*'s father," or, colloquially, "Everyone respects Pitt's father," "Pitt respects everyone's father," and "Everyone respects his own father." And now, in these quantified sentences, Frege says, the expression for the relevant concept *is* part of the content of the sentence.

The process of forming expressions for concepts may likewise be used to form expressions for functions, as we ordinarily conceive them, namely by starting with a complex singular term within which some simpler singular term occurs. And it may also be used to form expressions for relations between two objects, namely by removing from a sentence one or more occurrences of each of two different singular terms. This was of importance for second-order logic, admitting quantification over functions and relations.

Thus Frege's invention of the quantifier-variable notation yielded him several fundamental insights. First, the conception of concept-expressions as incomplete solved the problem of the unity of sentences and the thoughts they express. No glue is needed to make the parts of the sentence adhere to one another. The concept-expression or relation-expression is of its nature incapable of standing alone, but can be present only when its argument-places are filled by singular terms to form a sentence. Or else it is itself the argument of a quantifier, forming a different kind of sentence: it is made to adhere to such terms, or to have a quantifier attached to it, and cannot exist otherwise. Secondly, concept-formation does not consist solely of the psychological abstraction of some common feature from individual objects or of the process of applying Boolean operations to given concepts (conjoining or disjoining them). By the process of omitting singular terms from complete sentences, or, equivalently, of thinking of them as replaceable by other singular terms, we can arrive at expressions for concepts with new boundaries, and so at the concepts thus expressed. Moreover, such expressions were not, in general, actual *parts* of the sentences from which they were formed; they were, rather, patterns exemplified by different sentences. The expression "ξ respects ξ's father" occurs both in "Pitt respects Pitt's father" and in "Fox respects Fox's father"; the two sentences have, not just common words, but a common *pattern*. Thus we should not think of a sentence, or the thought it expresses, as formed out of its component parts, but of the components as attainable by analyzing the sentence; still less should we think of a concept-expression as formed out of its components, but as a result of analyzing

a sentence. Because apprehending the possibility of analyzing a sentence in a particular way requires us to see it as manifesting a certain pattern, which is not required for a simple grasp of the sentence's content, and because apprehending this possibility may be essential to recognizing the validity of some deductive inference, there is a creative ingredient in deductive inference. Such inference does not depend only upon a grasp of the contents of the sentences that figure in it; and this explains how deductive inference can lead us to new knowledge, which consideration of its role in mathematics makes evident that it does.

These ideas were expressed in *Begriffsschrift* and in *Grundlagen*. Part I of *Begriffsschrift* was devoted to sentential logic, and Part II to first-order predicate logic. Although Frege did not have the concept of the completeness of a logical system, he had in fact framed a complete formalization of first-order logic. Part III of *Begriffsschrift* is devoted to second-order predicate logic, involving quantification over concepts, relations, and functions; Frege never saw any reason for regarding the first-order fragment as especially significant. To explain second-order quantification in the same way as first-order quantification, Frege has to admit the notion of an expression for a concept of second level, formed by removing from a sentence one or more occurrences of an expression for a first-level concept, or of a relational or functional expression; these second-level concept-expressions all have different types of incompleteness. In Part III Frege gave his purely logical definition of a sequence; since previously the notion of an infinite sequence had usually been explained in temporal terms, as involving its successive construction step by step, or else a successive diversion of attention from one term to the next, Frege regarded this as an essential contribution to the program of expelling intuition from arithmetic in favor of purely logical notions. It was especially important for number theory, since the natural numbers themselves could be defined as the terms of a finite sequence beginning with 0 and proceeding from each term to its successor. Frege's definition of a sequence was so framed that, when the natural numbers are so defined, the principle of finite induction, sometimes claimed as a method of reasoning peculiar to arithmetic, becomes a direct consequence of the definition. Frege's definition of "sequence" is now generally known as the definition of the ancestral of a relation, namely the relation which the first term of a finite sequence has to the last term when each term but the last stands in the original relation to the next term. (It is named the "ancestral relation" because the relation "ancestor of" is the ancestral of the relation "parent of.")

There are three features of *Grundlagen* of especial interest to philosophy in general. The first is the distinction between the actual (*wirklich*) and the objective. Frege used "actual" to mean "concrete" in the sense in which concrete objects are distinguished from abstract ones; an object is actual if it is capable of affecting the senses, directly or indirectly. But something may be objective even though it is not actual; an example he gives is the Equator. You cannot see or trip over the Equator, but it is not fictitious or subjective; statements about it may be objectively true or false. We can make reference to objects which, though objective, are not actual, and make objectively true statements about them. Frege thus rejected what is now called "nominalism" as based on a fundamental error. This was crucial for his philosophy of arithmetic. He took numbers to be objects, objective but not actual: we can refer to them and make objectively true statements about them.

A principle greatly stressed in *Grundlagen* is what has come to be known as the "context principle": that it is only in the context of a sentence that a word has a meaning. It is noteworthy that the principle is formulated linguistically, as concerning the meanings of words in sentences, rather than, say, as "We can think of anything only in the course of thinking that something holds good of it." The interpretation of the dictum is contentious. At the very least, it is an assertion of the primacy of sentences in the order of explanation of meaning. We must first explain what, in general, constitutes the meaning of a sentence, and then explain the meanings of all small expressions as their contributions to the meanings of sentences in which they occur. When we look at how Frege applies the principle in the book, it appears to have a much stronger significance: namely that, to secure a meaning for an expression or type of expression, it suffices to determine the senses of all sentences in which it occurs. Frege never reiterated the context principle in any subsequent writing, although there is a strong echo of it in Part I of *Grundgesetze*.

The other salient feature is the first clear example of the linguistic turn, giving Frege a strong claim to be the grandfather of analytic philosophy. At a critical point of the book, Frege, having already argued that our notion of number is not derivable from sense-perception or intuition, asks, "How, then, are numbers given to us?" The question is both epistemological and ontological: how are we aware of numbers, and what guarantee is there that such objects as numbers exist? In answering it, Frege simply assumes that it can be equated to "How are meanings conferred on numerical terms?" He appeals immediately to the context principle; in virtue of this, the question reduces to, "What sense attaches to statements containing numerical terms?" A question about what objects exist and how we know of them is thus transformed into a question about the meanings of certain sentences.

However, those of Frege's ideas that most interested analytic philosophers when interest in his work revived were the ones he expounded in his middle period (1891–1906). Frege had no general term for meaning, in the sense in which the meaning of a word or expression comprises everything that a speaker must implicitly know about it in order to understand it. He distinguished three features which, in this sense, may contribute to the meaning of a word or sentence: force, tone, and sense. The force of an utterance is what distinguishes an assertion from a question, and Frege recognized only these two types of force: assertoric and interrogative. In English interrogative force is usually indicated by the inversion of the verb and subject; Frege insisted that the sense of a question inviting the answer "Yes" or "No" will coincide with that of the corresponding assertoric statement. What differentiates them is the significance of the utterance: in one case we ask whether the thought expressed is true, in the other we commit ourselves to its truth. It was important for Frege that only a complete utterance can carry force; a declarative sentence serving as, say, one clause of a disjunctive statement or as the antecedent of a conditional one does not have assertoric force, which is attached only to the statement as a whole. It was essential, Frege thought, not to construe the verb or predicate of a sentence as intrinsically containing the assertoric force within it. Natural languages usually lack any express means of indicating that assertoric force is to be attached to a sentence, but Frege considered this a defect of them. In his formal language he used a symbol for just this purpose, the "judgment-stroke" (often called by others the "assertion sign"). It is a philosophical mistake to speak of "judg-

ments" when all that we are concerned with is their contents; unless we are actually concerned with the act of recognizing them as true, we should speak in this connection of thoughts, rather than of judgments. Frege did not recognize imperatival or other kinds of force, though it may plausibly be argued that an imperative sentence expresses the thought that is true if the command is obeyed or the demand complied with. He simply declared that such a sentence expresses a command, not a thought.

What I have called "tone," and Frege called *Färbung* (coloring) is distinguished from sense in that it cannot affect the truth or falsity of what is said. The English sentences "He has died," "He is deceased," and "He has passed away" do not differ in sense, but only in tone. Likewise, where A and B are sentences, the complex sentences "A and B" and "Not only A but B" do not differ in sense, but in tone: if either is true, the other is true, even if it conveys an inappropriate suggestion. The sense of a whole sentence is the thought that it expresses; the sense of a part of a complex expression, including a sentence, is part of the sense of the whole.

In his middle period, Frege drew a distinction between the significance of an expression and what it signifies, which he had not done in his early period. For the thing signified, he confusingly chose the word "*Bedeutung*," the ordinary German word for "meaning": but the *Bedeutung* of an expression is not part of its meaning, where "meaning" is understood as specified above. It is not necessary, in order to understand a word or phrase, to know its *Bedeutung*, only its sense. Frege's term is conventionally rendered in English either "meaning" or "reference"; neither is happy. The *Bedeutung* of a singular term is the object we use the term to talk about. It is impossible just to know the *Bedeutung* of a singular term, even if that term is logically simple, i.e. it is a proper name in the restricted sense (Frege misleadingly called all singular terms "*Eigennamen*" – proper names). Frege followed Kant in holding that every object of which we are aware is given to us in a particular way; the sense of a singular term embodies the particular way in which its *Bedeutung* is given to us in virtue of our understanding of the term.

But it was not only singular terms which Frege took as having *Bedeutungen*: he ascribed them to every expression that could be a genuine constituent of a sentence, including incomplete ones such as concept-expressions and sentences themselves. He does not argue that any such expression must have a *Bedeutung*; he takes it for granted. The only question he canvasses is what kind of thing the *Bedeutung* of an expression of any given type should be taken to be. This causes much perplexity to those reading Frege for the first time: surely there is nothing to which a concept-expression or a sentence stands as a name stands to the object named. The only way to arrive at an understanding of Frege's notion of *Bedeutung* is to look at the use to which he puts it. That use is governed by four fundamental theses:

1 The *Bedeutung* of a part of a complex expression is not part of the *Bedeutung* of the whole.
2 But the *Bedeutung* of the whole depends uniquely upon the *Bedeutungen* of its parts.
3 If a part lacks *Bedeutung*, the whole lacks *Bedeutung*.
4 The *Bedeutung* of a sentence is its truth-value – its being true or its being false.

Thesis (1) follows from the fact that Sweden is not part of Stockholm, the capital of Sweden; and thesis (3) derives from the consideration that, if there is no such country

as Ruritania, then there is no such city as the capital of Ruritania. As for the identification of truth-values as the *Bedeutungen* of sentences in accordance with thesis (4), that follows from Frege's extensionalist logic, despite its failure to fit natural language. According to it, a subsentence of a complex sentence contributes to the truth-value of the whole solely by its own truth-value. Counterexamples from natural language then have to be explained away. Instances are sentences in indirect speech following verbs like "said that" and "believes that"; as is well known, Frege handled these by deeming the sentences following "that" to have a special, indirect sense, whereby their *Bedeutung* became the senses that they would express when in direct speech.

It is plain from these four theses that the *Bedeutung* of an expression constitutes its contribution to the determination of the truth-value of any sentence in which it occurs. This explains why Frege takes it for granted that any expression capable of occurring in a sentence without denying it a truth-value must have a *Bedeutung*. It also makes a Fregean theory of *Bedeutung* for a language equivalent to what we understand as a semantic theory for that language, which is a theory explaining how sentences of the language are determined as true or as false in accordance with their composition. The semantic value of an expression, in such a theory, is precisely that which contributes to the determination of the truth-value of a sentence in which that expression occurs. We may therefore equate the notion of the *Bedeutung* of an expression, as Frege conceived it, with that of its semantic value.

In a conventional semantic theory for a formalized language, the semantic value of an individual constant or other singular term is an element of the domain denoted by the term. The semantic value of a one-place predicate is a class of elements of the domain; the sentence resulting from putting the term in the argument-place of the predicate is true if the element denoted by the term is a member of the class constituting the semantic value of the predicate, false otherwise. Frege did not speak of the domain of quantification; so far as can be determined, he took the individual variables to range over all objects whatever, the *Bedeutung* of any term being such an object. Frege did not take the *Bedeutung* of a concept-expression to be a class. He called the *Bedeutung* of a concept-expression a "concept." This must not be understood in the sense in which we may speak of acquiring a concept or grasping a concept, which has to do with the senses expressed by words. In *Grundlagen*, the word "concept" (*Begriff*) had been used both in this way and in conformity with what was to become Frege's usage in his middle period; but, in that period, he took a concept to stand to a concept-expression as an object stands to a singular term, and thus not at all as the sense of that expression.

For Frege, a concept must be distinguished from a class, which was for him a particular kind of object. A class is the extension of a concept, comprising those objects that fall under the concept; but the extension of a concept is a derivative notion, only to be so explained. The relation of being a member of a class can be explained only as that of falling under a concept of which the class is the extension; any attempt to explain it in any other way turns the relation into that of part to whole, which is quite different. So we can attain the concept of a class only via that of a concept; and we can characterize any particular class only by citing a concept of which it is the extension.

Since a concept-expression was for Frege incomplete, so its *Bedeutung* cannot be an object of any kind, but must be likewise incomplete, an entity needing an object to

saturate it. This is a difficult conception, but is to be thought of by analogy with how we think of functions. We think of an arithmetical function as a principle according to which one number is arrived at, given another: not a *method* of arriving at the value, given the argument, but simply the association of the value to the argument. Its incompleteness consists in the fact that there is nothing to it save this association: its existence consists solely in its linking arguments to values. Likewise, the existence of a concept consists solely in its having certain objects falling under it, and others not falling under it. In facts, according to the doctrines of Frege's middle period, concepts simply are a particular type of function. For he regarded truth-values – truth and falsity – as themselves being objects. So a concept is a function which takes only truth-values as values, mapping an object falling under it on to the value *true*, and one not falling under it on to the value *false*. In the same way, a relation is a function with two arguments, all of whose values are truth-values.

In his early period (1874–85), Frege had not distinguished between significance and thing signified; he had used the one term "content" for both without differentiation. It was a great advance that in his middle period he sharply distinguished them as sense and *Bedeutung*. What we tacitly know in understanding a word or expression is its sense; its sense is the way its *Bedeutung* is given to us. It is not only of an object that it holds good that it must be given to us in a particular way: the same holds good of concepts, relations, and functions. For instance, an arithmetical function may be given to us by means of a particular procedure for computing its value, given its argument or arguments; other procedures might serve to determine the values of just the same function. For this reason, the content of any piece of knowledge that we may have concerning a given expression can never simply consist in our knowing its *Bedeutung*, but must be our knowing its *Bedeutung* as given in a particular way. The *Bedeutung* of an expression is therefore no part of its meaning, where this is what we grasp in understanding the expression: what we grasp is its *sense*. We may indeed grasp more than its sense, namely what was called above its tone. Sense is that part of the meaning of the expression that is relevant to the determination of a sentence containing it as true or as false. But the notions of sense and *Bedeutung* are closely connected: again, the sense of an expression is the way in which its *Bedeutung* is given to us. (The sense is *die Art des Gegebenseins*, usually clumsily translated "the mode of presentation.")

It is not only that each individual speaker must think of the *Bedeutung* of a word as given in some particular way, leaving it possible for different speakers to think of it as given in different ways. For successful communication, the speakers must know that the *Bedeutung* of a word, as each is using it, is the same. To ensure this, it must be a convention of the language that each associates with the word the same sense, that is, the same way of thinking of something as its *Bedeutung*. An imaginary example given by Frege in a letter to Jourdain is that the *Bedeutung* of the name "Afla" might be given as the mountain visible on the northern horizon from such-and-such a place, and that of the name "Ateb" as the mountain visible on the southern horizon from a certain other place. It may prove that the two names have as *Bedeutung* the very same mountain, which was not at first evident; the identity-statement "Afla and Ateb are the same" is informative and reports an empirical discovery. Famously, in his celebrated essay "Über *Sinn* und *Bedeutung*" of 1892, Frege used the example of the names "the Morning Star" and "the Evening Star," which both denote the planet Venus, to illustrate the

distinction between sense and *Bedeutung* and so explain how a true statement of identity could be informative (he had used the same example earlier in his lecture "Function and Concept" of 1891).

To know the sense of an expression is, therefore, to know how its *Bedeutung* is determined: not necessarily how *we* can determine it, since we may lack an effective means of doing so, but how, as it were, reality determines it in accordance with the sense we have given it. The sense of a part is part of the sense of the whole; the sense of any given expression is part of the sense of any more complex expression of which the given expression is part. So a grasp of the sense of an expression involves knowing how it may be put together with other expressions to form a complex expression – ultimately, a sentence – and how the sense of the complex is determined from the senses of its parts. To grasp the sense of a concept-expression is to apprehend a particular way of thinking of something incomplete as its *Bedeutung*, something that associates each arbitrary object with a truth-value: a concept that carries each object into the value *true* or the value *false* according as it falls under the concept or not. In general, we grasp the sense of a whole sentence by grasping the sense of each expression composing it, which is its contribution to the sense of the sentence as a whole; and to do this is to have a particular conception of the *Bedeutung* of each constituent, together with a grasp of how these *Bedeutungen* combine to yield the *Bedeutung* of each phrase and ultimately of the sentence itself. But the *Bedeutung* of a sentence is a truth-value; its sense Frege terms a *thought*. In the case of a sentence, the distinction between sense and *Bedeutung* is that between a thought and its truth-value. Thus to grasp a thought is to apprehend how it is determined – by reality, though not necessarily by us – as true or as false. And to grasp a sense that goes to compose a thought by being the sense of a constituent of a sentence that expresses that thought is to understand how the contribution to determining the truth-value of the thought that is made by that constituent is itself determined. In the words of *Grundgesetze*, Part I, the thought expressed by a sentence is the thought that the condition for its truth is fulfilled. This was an expression of what has become the most popular form of a theory of meaning, a truth-conditional theory: truth is the central notion of such a theory, and meaning is to be explained in terms of it.

Frege held that anyone who makes a judgment knows implicitly what truth and falsity are. We can express a thought without asserting or judging it to be true, which we do whenever we utter a sentence whose sense it is but to which assertoric force is not attached (e.g. when we ask whether it is true). When we judge the thought to be true, we "advance from the thought to the truth-value." But this advance is not a further thought, to the effect that the original thought is true; by prefacing the sentence expressing the thought with the words "It is true that", we do not confer assertoric force on it, but merely express the very same thought as before. That is why Frege says, in one of his posthumously published writings, that the word "true" seems to make the impossible possible. Frege held the notion of truth to be indefinable: he rejected the correspondence theory of truth, and any other such theory that professes to say what truth is.

Frege was vehemently opposed to psychologistic explanations of concepts, that is, of the senses of linguistic expressions. He opposed explanations in terms of the inner mental operations by which we acquire such concepts. The sense of any expression had

to be explained objectively, not subjectively, in terms of the conditions for the truth of sentences containing the expression. A thought, for Frege, is not one of the contents of the mind, as is a sense-impression or a mental image. These are subjective and incommunicable; but it is of the essence of thoughts to be communicable. Different people can grasp the very same thought; it cannot therefore be a content of any of their minds. This rejection of psychologism was of the greatest importance: it rescued the philosophy of thought and of language from explanations given in terms of private psychological processes. Frege's alternative explanation was neither so popular nor so successful. He recognized no intermediate category between the subjective and the wholly objective. He took thoughts and their component senses to constitute a "third realm": like the physical universe, its inhabitants are objective, but, unlike it, they are not in time or space or perceived by the senses. But it is only through our grasp of the inhabitants of the third realm that mere sense-impressions are converted into perceptions, and so we become aware of the external world. We can grasp thoughts and express them: but we human beings can grasp them only as expressed in language or in symbolism.

Frege's attitude to language was ambivalent. He viewed natural language as full of defects: only when it was conducted by means of a purified language, such as his logical symbolism, could deductive reasoning be confidently relied on. So some of the time he inveighs against language, declaring that philosophy must struggle against it and that his real concern is with thoughts and not with the means of their expression. Yet a great deal of his discussions are concerned precisely with language and its workings. His philosophical logic is not a theory of thought, independent of language: it is a systematic theory of meaning, applicable directly to a language purified of the defects of our everyday speech, but indirectly to natural language. The power of his theory of meaning rests upon the capacity of predicate logic – the logic he first invented – to analyze the structure of a great range of sentences and of the thoughts they express. Although many of his ideas were not found acceptable by later analytic philosophers, his theories were seen as a better model of what philosophy should aim at, in framing its basic theories of meaning and of thought, than anything supplied by any other philosopher; and his discussions of problems within that realm a better place to start from than any other.

In recent years there has been a great revival of interest in Frege's philosophy of mathematics, the late George Boolos being one of those to have contributed greatly to it. The comparison between Frege's *Die Grundlagen der Arithmetik* and Richard Dedekind's *Was sind und was sollen die Zahlen?*, two books which approach the same subject matter very differently, is extremely fruitful. Dedekind is concerned to characterize the abstract structure of the sequence of natural numbers; having done so, he arrives at that specific sequence by an operation of psychological abstraction, a quite illegitimate device much favored by mathematicians and philosophers of the time. He acknowledges the use of the natural numbers to give the cardinality of finite classes, but only as a minor corollary. For Frege, by contrast, that use is central. It was for him the primary application of the natural numbers, and must therefore figure in their definition. "It is applicability alone," he wrote in Part III of *Grundgesetze*, "that raises arithmetic from the rank of a game to that of a science." He strongly opposed appeal, such as that made by J. S. Mill, to empirical notions having to do with one or other par-

ticular type of application, in defining the natural numbers or the real numbers. But he thought it essential that, in defining them, the general principle underlying all their applications should be made central to their definition. Hence natural numbers were to be presented as finite cardinals: the operator in terms of which all numerical terms were to be framed was "the number of *x*'s such that . . . *x* . . . ," where of course the gap was to be filled by an expression for a concept of first level.

Frege's aim was to show that arithmetical were derivable from purely logical principles. A description of physical space as non-Euclidean is intelligible; so Euclidean geometry is not analytically true. By contrast, any attempt to describe a world in which the truths of arithmetic fail is incoherent. Since Frege characterizes logical notions as those which are topic-neutral, applying to things of every kind, arithmetical notions are already logical ones. But, like Russell's "axiom of infinity," a proposition may be expressed in logical terms without its truth being guaranteed by logic. It therefore remains to be shown that what we take to be the fundamental truths of number theory are derivable from purely logical principles.

Frege endorsed the definition of equicardinality that was becoming generally accepted by mathematicians, in particular Cantor:

> There are just as many *F*s as *G*s iff there is a relation which maps the *F*s one-to-one on to the *G*s.

If there is a cup on every saucer on the table, and every cup on the table is on a saucer, we shall know that there are just as many cups as saucers on the table without necessarily knowing how many of each there are. In *Grundlagen* Frege enunciates a basic principle governing his cardinality operator:

(*) The number of *F*s = the number of *G*s iff there are just as many *F*s as *G*s,

"just as many as" being interpreted in accordance with the foregoing definition. He decides that the cardinality operator cannot be defined contextually, but requires an explicit definition: the one that he chooses is:

> The number of *F*s = the class of concepts *G* such that there are just as many *F*s as *G*s.

Here Frege appeals to the notion of a class for the first time, although he never again considers classes of concepts rather than of objects. But the appeal is solely for the purpose of framing an explicit definition of the cardinality operator; Frege uses it for nothing else than proving the principle (*) from it: all the theorems he goes on to prove about the natural numbers are derived from (*) alone, without further recourse to the definition of "the number of."

The theory sketched in *Grundlagen* is elaborated and fully formalized in *Grundgesetze*, Part II. *Grundgesetze* makes extensive use of the notion of classes, or, rather, of Frege's generalization of it, that of value-ranges: a class is the extension of a first-level concept, while a value-range is the extension of a first-level function of one argument. The latter notion is for Frege the more fundamental one, since concepts are for him a special kind of function. Frege had convinced himself that the notion of a value-range was a logical one. The Basic Law governing the operator forming terms for value-ranges is Law V:

the value-range of f = the value-range of g iff $f(x) = g(x)$ for every x.

It was of course this law which gave rise to the contradiction. Because of this, interest has centered upon a possible modification of Frege's construction of number theory, in which there are no value-ranges or classes, but the cardinality operator, governed by (*), is treated as primitive. Attention has focused on what is now called "Frege's Theorem," namely the proposition that, using Frege's definitions of "0," "successor," and "natural number," all of Peano's axioms, and hence the whole of second-order Peano arithmetic, can be derived in a second-order system from (*) alone. Opinions vary about how close this result brings us to Frege's goal of proving the truths of number theory to be analytic.

While most attention has been paid to Frege's foundations for number theory, some has been given to his foundation for the theory of real numbers, expounded in the incomplete Part III of *Grundgesetze*. Unlike both Cantor and Dedekind, Frege does not first construct the rational numbers and then define real numbers in terms of them: true to his principle that types of number are distinguished by their applications, and holding that both rationals and irrationals serve to give the magnitude of a quantity, he simply treats rational numbers as a kind of real number, defining the latter directly. While cardinal numbers answer questions of the form "How many . . . ?," real numbers answer those of the form "How much . . . ?" Any such question that can be answered by a rational number can also be answered by an irrational number. There are various quantitative domains – lengths, durations, masses, electric charge, etc.; within each, the magnitude of a quantity is given as the ratio of the given quantity to some chosen unit quantity; these ratios are the same from domain to domain. Thus real numbers are to be defined as ratios of quantities belonging to the same domain; such a definition accords with Frege's general tenet, that the definition of a type of number should incorporate the general principle underlying all its applications.

In the sections of Part III included in Volume II of *Grundgesetze*, Frege is concerned to characterize quantitative domains, and he identifies them as groups of permutations of an underlying set satisfying certain conditions. Unknown to Frege, this work had been partially anticipated by Otto Hölder in an article of 1901. Neither Frege nor Hölder uses explicit group-theoretical terminology. Both of them were concerned with groups with an ordering upon them. Hölder is generally credited with having proved the Archimedean law from the completeness of the ordering, which Frege also proved; but Frege's assumptions are much weaker than Hölder's. Frege assumes only that the ordering is right-invariant and that it is upper semi-linear (the ordering is linear upon the elements greater than any given element); Hölder makes the further assumptions that it is also left-invariant, fully linear, and dense. This preliminary part of Frege's construction of the foundations of analysis contains substantial contributions to group theory, and Part III as a whole presents pregnant ideas about how real numbers should be explained.

Frege's work on the philosophy of mathematics offered an explanation of how deductive reasoning can extend our knowledge, and a conception of the significance of the applications of a theory to its foundations. It also challenges us to say on what our recognition of mathematical truth rests, if not on pure logic or, more generally, on purely conceptual truths. But it offers another challenge not so often recognized.

Frege's attempt in Part I of *Grundgesetze* to justify his introduction of value-ranges was undoubtedly a failure: he was attempting simultaneously to specify the domain of his individual variables and to interpret his primitive symbols over that domain. But he was facing a problem that is usually left untackled: how can we without circularity justify the existence of domains sufficiently large to contain the objects of our fundamental mathematical theories such as number theory and analysis? Until a convincing answer is given to this question, we shall not have a satisfying philosophy of mathematics.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

2

Bertrand Russell (1872–1970)

THOMAS BALDWIN

Russell was the most important British philosopher of the twentieth century. At the start of the century he helped to develop the new theories that transformed the study of logic at this time, but his greatest contribution was not to logic itself. Instead it lay in developing and demonstrating the philosophical importance of this new logic and thereby creating his "logical-analytic method," which is the basis of the analytical style of philosophy as we know it today. The result is that we can still read Russell's writings as contributions to contemporary debates. He is not yet someone whose works belong only to the history of philosophy. He lived to be nearly 100 and there is every reason to expect that some of his writings will have an active life and age at least as great as his. A classic instance is provided by his introduction to philosophy, *The Problems of Philosophy* (his "shilling shocker" as he liked to call it), which, though published in 1912, remains one of the best popular introductions to the subject.

Early life

Despite the fact that as a philosopher Russell remains almost a contemporary, in other respects his life now seems very distant from us. His family, the Russells, was one of the great Liberal families of British politics: his paternal grandfather, Lord John Russell, had been Prime Minister twice during the first half of the nineteenth century, and Russell describes meeting Mr. Gladstone several times. His parents, Viscount Amberley and his wife Kate, were friends with John Stuart Mill, who agreed to act as an honorary godfather to their young son Bertrand. Mill in fact died during the following year, too soon for Bertrand to make his acquaintance. Much more traumatic for the young child, however, was the death of both his parents soon afterwards, so that in 1876 he was left at the age of 4 in the care of his grandparents, indeed of just his grandmother after his grandfather's death in 1878. Russell described his lonely childhood in his *Autobiography*. His rebellious elder brother Frank was sent away to school, but Bertrand ("a solemn little boy in a blue velvet suit," 1967: 30) was educated at home, brought up in a constricting atmosphere whose narrow limits were fixed by his grandmother's strict Presbyterian beliefs. The only refuge that the young Bertrand found was in the privacy of his own thoughts; he kept a secret diary in code in which he set down his growing doubts about religious orthodoxy. There can be little doubt that Russell's

troubled later emotional life (he had four marriages) was affected by this childhood. In his writings there are many passages in which he refers indirectly to it; for example, when writing in 1916 about the difficulties of marriage, he remarks that "The fundamental loneliness into which we are born remains untouched, and the hunger for inner companionship remains unappeased" (1916: 191).

From an early age his precocious talent in mathematics had been recognized, and in 1890 he went to Trinity College, Cambridge to study mathematics. Despite his delight in escaping from his grandmother, however, he soon found himself dissatisfied with the antiquated teaching of mathematics at Cambridge. So in 1893 he switched to the study of philosophy, a subject into which he had been initiated through membership of the Cambridge "Apostles" (a private society largely dedicated to the discussion of philosophy) and friendship with the philosopher J. M. E. McTaggart, then a young Fellow of Trinity. In 1894 he obtained a first class result in his final examinations and almost immediately started work on a dissertation in the hope of winning a prize fellowship at Trinity College. At the same time he married, and then travelled with his first wife, Alys, to Germany. In his *Autobiography* he recounts a moment of clear-minded future resolution during this honeymoon:

> During this time my intellectual ambitions were taking shape. I resolved not to adopt a profession, but to devote myself to writing. I remember a cold, bright day in early spring when I walked by myself in the Tiergarten, and made projects of future work. I thought that I would write one series of books on the philosophy of the sciences from pure mathematics to physiology, and another series of books on social questions. I hoped that the two series might ultimately meet in a synthesis at once scientific and practical. My scheme was largely inspired by Hegelian ideas. Nevertheless, I have to some extent followed it in later years, as much at any rate as could have been expected. The moment was an important and formative one as regards my purposes. (1967: 125)

This passage (though written with the benefit of hindsight) is remarkably prophetic; Russell had no settled profession (he held teaching positions for only about ten years) and for most of his life he made his living by writing, in which he had remarkable proficiency. He wrote about seventy books, which do indeed form two series: there are about twenty books of philosophy, mostly on "the philosophy of the sciences"; and many of the rest concern "social questions," though it cannot be said that the two series meet in a synthesis. His most popular book was *A History of Western Philosophy*, despite the fact that this is an unreliable and distinctly idiosyncratic book in which Russell devotes most space to ancient and medieval philosophy.

Breaking with idealism

During the 1890s the dominant school of philosophy in Cambridge, as in Britain generally, was idealist. Under McTaggart's influence Russell chose to work within a broadly idealist framework (as the Tiergarten testament quoted above shows), and the project he selected for his fellowship dissertation was that of providing a revised a priori foundation for geometry, one that would take account of the possibility of non-Euclidean geometries in a way that Kant's famous account does not. Russell argued that Kant's

conception of the conditions of the possibility of experience had been too limited. What is important, and thus a priori, is that space be of a constant curvature, but it is not an a priori matter just what its curvature is – for example whether it is zero, as Euclid maintained, or positive, as Riemann proposed. Russell was duly elected to a six-year prize fellowship at Trinity College in 1895 and in 1897 he published a revised version of the dissertation, *An Essay on the Foundations of Geometry*.

As well as offering a Kantian foundation for geometry Russell argued in a Hegelian fashion that within the abstract conception of points in space characteristic of geometry there are "contradictions," which can only be resolved by incorporating geometry into an account of the physical structure of space. This led him into a study of the foundations of physics and in particular to a study of the problems associated with the continuity of space and time. He initially approached these matters with the presumption that these problems arise from a conflict between, on the one hand, the fact that individual points or instants differ only in respect of their relations and, on the other, the requirement that "all relations are internal," which he took to imply that differences in the relationships between things are dependent upon other differences between these things. Russell summed up the conflict here as "the contradiction of relativity": "the contradiction of a difference between two terms, without a difference in the conceptions applicable to them" ("Analysis of Mathematical Reasoning," *Papers*, 2: 166). The presumption that there is a contradiction here was a commonplace among the idealist logicians of the period, such as F. H. Bradley, and was central to their denial that there are any relational truths and thus to their metaphysical monism. It was therefore by calling this presumption into question that Russell made his break with idealism. The key to this was his affirmation of the independent reality of relations, which he proposed in his 1899 paper "The Classification of Relations" (see *Papers*, 2). Once this move was made, the alleged "contradiction of relativity" is dissipated and Russell was free to approach the issues raised by the continuity of space and time afresh.

Although Russell's papers from this period show him finding his own way to this anti-idealist thesis, he always acknowledged the decisive importance of G. E. Moore's writings at this time ("It was towards the end of 1898 that Moore and I rebelled against both Kant and Hegel. Moore led the way, but I followed closely in his footsteps," 1995a: 42). G. E. Moore was two years younger than Russell. Having been drawn from the study of classics to that of philosophy at Trinity College partly through Russell's influence he graduated in 1896 and completed his own, successful, dissertation for a prize fellowship in 1898. It is in this dissertation that Moore works out his own break with idealism. His basic claim is that of the unqualified reality of the objects of thought, propositions, as things is in no way dependent upon being thought about. Moore further maintained that there is no reason to duplicate ontological structures by hypothesizing the existence of facts for true propositions to correspond to. Instead there are just propositions and their constituents: the world just comprises the totality of true propositions, and an account of the structure of propositions is an account of the structure of reality itself. One implication of this is that the structure of space is independent of our experience of it and, therefore, of the conditions under which experience of it is possible. So Moore was very critical of Russell's neo-Kantian account of geometry, and Russell quickly came to agree with Moore on this matter (see MOORE).

Russell read Moore's dissertation at an early stage, and immediately accepted many of Moore's central points, including in particular his conception of a proposition. As we shall see, many of his later difficulties can be traced back to this. But at the time Russell was exhilarated by the possibilities that this new realist philosophy opened out before him:

> But it was not only these rather dry, logical doctrines [concerning the reality of relations] that made me rejoice in the new philosophy. I felt it, in fact, as a great liberation, as if I had escaped from a hot-house on to a wind-swept headland. (1995a: 48)

The principles of mathematics

This tremendous sense of intellectual liberation quickly became focused on a new project, which was to dominate Russell's thought and life for the next ten years: the "logicist" project of demonstrating that "all mathematics is Symbolic Logic." The occasion which fired Russell's enthusiasm for undertaking this project was his visit in July 1900 to the International Congress of Philosophy in Paris, where he heard Peano discuss his formalization of arithmetic using new logical techniques. Peano did not himself seek to provide a purely logical foundation for mathematics; he did not offer logical definitions of the concepts "0," "successor," and "number" which occur in his postulates. But, on hearing him, Russell jumped to the hypothesis that definitions of this kind should be possible, and thus that mathematics is, in the end, just logic. In making this jump Russell was drawing on his recent close study of the philosophy of Leibniz, which, fortuitously, he had undertaken the year before (simply because he stood in for McTaggart who should have been teaching it). Russell recognized that Leibniz had also conceived this hypothesis but had been prevented from demonstrating it, largely because of the inadequacies of the traditional logic to which he adhered. But with the richer resources of the logic employed by Peano (which Russell immediately used to develop a new logic of relations), Russell supposed that Leibniz's logicist hypothesis could now be vindicated.

An important aspect of Russell's new project was the opportunity it provided him to continue his criticisms of idealist philosophy:

> The questions of chief importance to us, as regards the Kantian theory, are two, namely, (1) are the reasonings in mathematics in any way different from those of Formal Logic? (2) are there any contradictions in the notions of time and space? If these two pillars of the Kantian edifice can be pulled down, we shall have successfully played the part of Samson towards his disciples. (1903: 457)

In providing a negative answer to the second of these questions Russell drew on the work of the great German mathematicians of the nineteenth century, Dedekind, Weierstrass, and Cantor, whose work he had discovered a few years earlier but had not then appreciated fully because of his attachment to idealist doctrines. He now devoted a central section of his new book, *The Principles of Mathematics*, to a careful exposition of their philosophy of the infinite, from which he concluded "that all the usual arguments, both as to infinity and as to continuity, are fallacious, and that no definite

contradiction can be proved concerning either" (1903: 368). Since the idealist logicians had advanced the opposite view, and then used the infinity of space and time to argue for their unreality, Russell felt that he was here providing a definitive refutation of their position.

In laying out this new philosophy Russell worked at extraordinary speed. He started writing *The Principles of Mathematics* in October 1900 and by the end of the year he had completed a first draft: the book as we now have it is more than 500 pages long and 300 of these come unchanged from that first draft. This period was, he wrote,

> an intellectual honeymoon such as I have never experienced before or since. Every day I found myself understanding something that I had not understood on the previous day. I thought all difficulties were solved, all problems were at an end. (1995a: 56)

But, he continues,

> The honeymoon could not last, and early in the following year intellectual sorrow descended upon me in full measure. (1995a: 56)

The main reason for the onset of this sorrow was his discovery, early in 1901, of "the contradiction," now usually known as "Russell's paradox." This is a contradiction that can be easily demonstrated just at the point at which one seeks to develop elementary logic into set theory in order to show how arithmetic can be established on the basis of logic alone. Russell discovered the contradiction when reflecting upon Cantor's "paradox" that there is no greatest cardinal number. Cantor's paradox rests on the theorem that the number of subsets of a given set S is always greater than the number of members of S itself. Cantor proves this theorem by deducing a contradiction from the hypothesis that there is a one-to-one correlation between the subsets of S and the members of S, which would imply, on the contrary, that their numbers are the same. The contradiction arises as follows: consider that subset of S whose members are just those members of S which do not belong to the subset of S with which they are correlated under the hypothesized correlation. Since this so-called "diagonal" set D is a subset of S it too must be correlated with a member of S, say d. Cantor now asks whether d belongs to D: given the way d and D have been defined, it turns out that d belongs to D if and only if d does not belong to D, from which it is easy to derive the explicit contradiction that d both belongs to D and does not belong to D.

The step from Cantor's theorem to Russell's paradox is very simple. Instead of Cantor's hypothetical one-to-one correlation between the members of a set and its subsets, consider instead the non-hypothetical identity relation between anything and itself. Then the analogue of Cantor's "diagonal" set D of things that are not members of the set with which they are hypothetically correlated becomes simply the set R of things that are not members of themselves. Under the "identity" correlation R is of course "correlated" with itself; hence in asking whether R belongs to that with which it is correlated we are simply asking whether R belongs to R. But since R just is the set of things which do not belong to themselves, it follows that R belongs to R if and only if R does not belong to R. This too immediately gives rise to an explicit contradiction, but in this case the derivation does not depend on a hypothetical

correlation which is thereby proved not to exist, but only on the non-hypothetical identity of a thing with itself which cannot be rejected. So in this case there is no obvious positive conclusion to be drawn comparable to Cantor's theorem. There is instead the utterly dismaying implication that there is something seriously amiss in the foundations of logic.

As soon as he had discovered this contradiction Russell communicated it to Frege, whose works he had just recently read properly for the first time and recognized for what they were, namely much the most sophisticated attempt to develop a logicist program of the kind he was also engaged upon. Frege received Russell's letter just as the second volume of his *Grundgesetze* was in press, and added the famous Appendix II, which begins, "Hardly anything more unfortunate can befall a scientific writer than to have one of the foundations of his edifice shaken after the work is finished." Frege did in fact suggest a way of circumventing Russell's paradox, but he could not show that it worked, and it is now known not to. Russell also attempted to find a way around the paradox and ended *The Principles of Mathematics* with a tentative proposal that the definition of the set *R* is ill-formed because a set, being of a different "type" from that of its members, cannot be a member of itself. As we shall see below, this is a proposal to which he returned later; but in the context of the *The Principles of Mathematics* it was not easy for him to advance it, since it conflicts with the conception of logic advanced there, namely that the truths of logic are truths which are absolutely universal, and are therefore not to be restricted by considerations arising from the type of thing under discussion.

The contradiction was not the only problem to delay publication of *The Principles of Mathematics* until 1903 and to dominate his research for the next few years. He also ran into a tangle of difficulties concerning the structure of judgment, with which he continued to grapple thereafter. These difficulties are, broadly, of two kinds: concerning (1) the unity of judgment and (2) the structure of general judgments.

In *The Principles of Mathematics* Russell's treatment of these matters is expressed through a discussion of the structure of propositions, which, following Moore, he takes to comprise both the objects of judgment and the objective structure of the world. So conceived, propositions are not representations which, when true, correspond to a fact. Instead true propositions just are facts – there is no difference between the death of Caesar and the (true) proposition that Caesar is dead. Since his death is something that befell Caesar himself, Caesar is himself a "constituent" of the proposition that Caesar is dead. Indeed the proposition just is a "complex" whose constituents are Caesar and death (which is a "predicate").

The difficulty that now arises concerning the "unity" of judgment is that of explaining how it is that a complete proposition is constituted. Russell discusses this in connection with the proposition that A differs from B. The constituents here are A, B, and difference; but specifying them does not yet specify the proposition in question, for they are also the constituents of the different proposition that B differs from A (because difference is a symmetric relation this is actually a poor case to have taken. The point is much clearer with an asymmetric relation, such as occurs in the proposition that A is larger than B, which is manifestly different from the proposition that B is larger than A although it has the same constituents). Russell sums up his discussion:

> The difference which occurs in the proposition actually relates A and B, whereas the difference after analysis is a notion which has no connection with A and B. (1903: 49)

Furthermore, Russell notes, it does not help if one adds that the difference in the case we want is a difference *of* A *from* B, for all that these additions do is to add further relations to the supposed constituents of the proposition without "actually relating" the relation of difference to the right terms. The problem is that

> A proposition, in fact, is essentially a unity, and when analysis has destroyed the unity, no enumeration of constituents will restore the proposition. (1903: 50)

The point raised here is one that was to come back to plague him. In *The Principles of Mathematics* Russell, having identified it, simply sets it aside for further treatment. This is disappointing, for here there is a straightforward challenge to the Moore–Russell conception of a proposition as a "complex whole" comprised of its elementary constituents. Indeed the point was not new: this difficulty concerning relational judgments had been famously set out by F. H. Bradley in chapter III of *Appearance and Reality* (1893). The person who first saw clearly the way to defuse the issue here was Frege: for through his famous "context" principle (presented in his *Grundlagen*, 1884) that it is only in the context of a sentence that a word has a meaning, he acknowledges the irreducible primacy of judgments instead of regarding them as "complexes" to be constructed out of elementary constituents. We shall see below that Russell's theory of descriptions includes a partial acknowledgment of the context principle; but he himself never generalizes it to solve this problem of the unity of judgment.

The other general type of difficulty that Russell encountered in *The Principles of Mathematics* concerns the structure of general propositions such as the proposition that I met a man. The difficulty here is supposed to come from the fact that, on the one hand, the concept *a man* is a constituent of this proposition; but, on the other hand, such a concept

> does not walk the streets, but lives in the shadowy limbo of the logic-books. What I met was a thing, not a concept, an actual man with a tailor and a bank-account or a public house and a drunken wife. (1903: 53)

Russell's argument here is intuitive and questionable. But it is clear that the tension arises from the dual role of propositions as both objects of thought (and thus constituted of concepts such as *a man*) and situations within the world (and therefore constituted not from general concepts, but from particular men).

Russell's way of resolving this tension is to say that concepts such as *a man* "denote" the things that a proposition in which they occur is "about"; and it is the things that are denoted in this way that "walk the streets" etc. It is difficult at first not to interpret this talk of that which a proposition is "about" as a way of implicitly specifying a "truth-maker" for the proposition distinct from the proposition itself; but this of course would be entirely inimical to Russell's conception of a proposition. Furthermore, Russell develops his account of denoting in such a way as to make this interpretation inappropriate. For he goes on to argue that

> *some man* must not be regarded as actually denoting Smith and actually denoting Brown and so on: the whole procession of human beings throughout the ages is always relevant to every proposition in which *some man* occurs, and what is denoted is essentially not each separate man, but a kind of combination of all men. (1903: 62)

As Russell acknowledges, such a "combination of all men" is a very paradoxical object; in the case of *a man*, what is denoted is supposed to be Smith or Brown or . . . (for the whole human race).

It is in fact clear enough what Russell is seeking to do here: namely, to effect a reduction of general propositions to propositions that involve only disjunctions or conjunctions of singular propositions, for which the tension between propositions as objects of thought and as truthmakers is not so acute. He writes that "the notion of denoting may be obtained by a kind of logical genesis from subject-predicate propositions" (1903: 54), and as Geach has observed, Russell's account is in this respect comparable to medieval theories of *suppositio*. But Russell muddles things by supposing that he needs to hold that there are disjunctive and conjunctive combinations of things denoted by the denoting concepts that occur in the general propositions. These are the "paradoxical objects" which the propositions in question are to be "about." Not only are such objects intrinsically objectionable (as Russell himself acknowledges, 1903: 55n.), this way of coming at the matter makes it impossible to provide a coherent treatment of propositions involving multiple generality, since there is no way of representing scope distinctions, such as the distinction between the two ways of interpreting "Everyone loves someone."

Russell in effect acknowledges this point himself when discussing variables, both free and bound. He would like to handle free variables within his theory of denoting as cases of the denoting concept *any term*; but he can see that this approach does not deal properly with the role of repeated variables. His discussion of bound variables occurs as part of his exposition of Peano's conception of "formal implication" as a universally quantified conditional, as in "for all x, if x is a man then x is mortal." Russell wants to be able to apply his theory of denoting concepts to the quantifier "for all x," as a concept denoting some combination of things which the whole proposition is "about." But he can also see that where there are multiple quantifiers binding different variables (as in both interpretations of "Everyone loves someone") the variables are tied to the quantifiers in a way that blocks off this conception of the denotation of a quantifier. So although he cannot bring himself to say so explicitly, his theory of denoting concepts is inadequate to the new logic of quantifiers and variables upon which the logicist project of *The Principles of Mathematics* is founded.

The theory of descriptions

Two years later, in 1905, Russell published his most famous paper, "On Denoting." He begins by, in effect, developing his earlier discussion of formal implication into a systematic account of the propositions expressed by sentences involving what he now calls "denoting phrases" such as "all men," "a man," and "no man." There is now no talk of denoting concepts; instead he uses the universal quantifier and bound variables to specify the propositions expressed by these sentences by reference to the truth of simpler propositions. Thus he now says that the proposition expressed by "I met a man" is the

proposition that propositions of the type "I met x, and x is human" are not always false (*Papers*, 4: 416).

Clearly, much is here assumed, for example the interdefinability of the existential and universal quantifiers. There is also a degree of oversimplification, since, as he recognizes when dealing with multiple quantifiers, he actually needs to specify the variable in the quantifier; in the case above he should have said "are not always false of x." Setting these points aside, what is worth considering is whether Russell offers any general account of quantifiers and variables to replace the theory of denoting concepts that has been tacitly discarded. In "On Denoting" itself Russell is unhelpful: he just says "Here the notion '$C(x)$ is always true' is taken as ultimate and indefinable" (*Papers*, 4: 416). If, however, one looks ahead to the discussion of the universal quantifier in the introduction that Russell wrote to *Principia Mathematica* (1910), one finds him using the language of "ambiguous denotation" to sketch what is now recognizable as a substitutional account of the quantifier. For he now says that we assert a universal proposition in order to condense the assertion of all the substitution instances "ambiguously denoted" (1910: 40) by the propositional function that occurs in our universal proposition; and the truth of the universal proposition depends on the "elementary truth" of all these substitution instances (p. 42). Since the motivation behind the original theory of denoting concepts was that propositions involving all men, a man, etc. are in some way "about" their instances, it is unsurprising that he ends up with a substitutional treatment of quantification, dealing, of course, with substitutions in propositions, not sentences.

Russell's main topic in "On Denoting" is the structure of propositions whose expression involves definite descriptions, phrases such as "The present President of the USA." In *The Principles of Mathematics* he had applied his theory of denoting concepts to such propositions in order to provide an account of why it is that true identities, propositions expressed by sentences such as "Bill Clinton is the present President of the USA," are of interest to us (1903: 62–4). The problem here is familiar: if we just take it that we have two names for the same thing, and that the proposition's constituents are just the thing thus named twice and the relation of identity, it seems that the very same proposition is also expressed by "Bill Clinton is Bill Clinton," which is of no interest to us. In *The Principles of Mathematics* Russell took it that this problem is solved by the hypothesis that the description "the present President of the USA" introduces a corresponding denoting concept into the proposition expressed through its use, which of course does not occur as a constituent of the proposition expressed by "Bill Clinton is Bill Clinton."

Russell does not explain how this denoting concept occurs as a constituent of the proposition expressed, and he says other things about it on the basis of which it is easy to reopen the old problem. Since the proposition cannot comprise Bill Clinton's identity with the denoting concept in question, it seems that it must comprise Bill Clinton's identity with the thing denoted by the denoting concept, that which the proposition is "about" in Russell's intuitive sense. But this thing is of course just Bill Clinton himself, and we have now come back to the difficulty of showing why this proposition is of any interest to us since it is equally expressed by "Bill Clinton is Bill Clinton." If one looks to Russell's extensive unpublished writings on this matter from the period 1903–5 (see *Papers*, 4), one can, I think, see him identifying this difficulty for his old position. He also begins to think about a different issue, also problematic for his old position, which

arises from his critical reaction to some works by the contemporary Austrian philosopher Alexius Meinong, which he studied closely at this time.

In this case the issue concerns the proper treatment of "empty" descriptions, descriptions such as "the present King of France" which, though meaningful, describe nothing that actually exists. Russell interpreted Meinong as advancing a theory of "objects" according to which empty descriptions of all kinds describe an object, even descriptions of impossible objects such as "the round square." In "On Denoting" and thereafter Russell ridiculed this position as conflicting with "the robust sense of reality" which "ought to be preserved even in the most abstract studies" (1919: 169–70); but in truth Meinong's position was a good deal more subtle than Russell appreciated. What is important here, however, is that through thinking about Meinong's work Russell came to appreciate the importance of constructing a theory that would allow for the possibility of meaningful sentences that include empty descriptions – sentences such as "The present King of France is bald." Since Russell took it that the meaning of a name was just the name's bearer, it followed at once that such descriptions were not names. This, he recognized, did not rule out the treatment of such descriptions as introducing denoting concepts into a proposition. But, he argued, there was still a problem: the proposition expressed by "The present King of France is bald" ought to be *about* the present King of France. But there is no such thing; so the theory implies that the proposition is about nothing, which Russell takes to imply that "it ought to be nonsense." But, he continues, "it is not nonsense, since it is plainly false" ("On Denoting," *Papers*, 4: 419).

This argument is condensed, but the way to understand the crucial move from being "*about* nothing" to "being nonsense" is to connect the "*aboutness*" thesis with the thesis that the truth or falsehood of a proposition with a denoting concept depends on the truth or falsehood of the propositions specified by reference to that which the first proposition is *about*. For in the light of this, a proposition *about* nothing will be one for which there are no such propositions to determine its truth or falsehood. So it can be neither true nor false itself; but this is absurd since propositions are inherently either true or false. Hence it follows that the sentence with an empty description fails after all to express a proposition: in which case "it ought to be nonsense."

We now have the two "puzzles" which, Russell says in "On Denoting," an adequate theory of descriptions must solve: (1) Why is it that questions about identity are often of interest to us – how can it be that George IV wished to know whether Scott was the author of *Waverley* but did not wish to know that Scott was Scott? (2) What propositions are expressed by sentences with empty descriptions, such as "The present King of France is bald," and how is their truth or falsehood determined? Before showing how his new theory can solve these puzzles, Russell explains why he rejects the view that one should account for these puzzles by distinguishing between the "meaning" and the "denotation" of a description. He specifically mentions Frege in connection with this view and it is natural to think of him as arguing here against Frege's theory of *Sinn* and *Bedeutung* which was of course precisely introduced to handle the first of Russell's puzzles and seems well suited to handle the second (see FREGE).

Russell's argument in "On Denoting" against the "Fregean" position is notoriously obscure. His conclusion is that there is an "inextricable tangle" in the account this position offers of the relation between the meaning and the denotation of a description

(*Papers*, 4: 422); but in truth it is Russell's own discussion which certainly appears to be an inextricable tangle. My own view is that Russell's argument is vitiated by the fact that he adapts Frege's theory to his own different conception of a proposition, and thereby crucially distorts Frege's actual position. (Russell himself acknowledges the differences here: *Papers*, 4: 419 n. 9.) This can be illustrated by considering a slightly later, and considerably clearer, discussion by Russell of the same issue, in "Knowledge by Acquaintance and Knowledge by Description," in which he is considering the structure of the proposition expressed by "the author of *Waverley* is the author of *Marmion*" (*Papers*, 6: 159). Russell argues here that (1) the proposition involves an identity; but (2) plainly does not involve the identity of the meanings of "the author of *Waverley*" and "the author of *Marmion*"; so (3) it must comprise the identity of the denotations determined by the meanings of "the author of *Waverley*" and "the author of *Marmion*." But these denotations are just Scott himself, so the proposition in question is just the proposition that Scott is Scott. Yet the whole point of the meaning/denotation theory was to preserve the distinction between interesting and trivial identities.

For Frege this argument simply gets off on the wrong foot. The thought expressed by "the author of *Waverley* is the author of *Marmion*" is indeed a thought about identity; but the thought itself is not structured by the relation of identity, but by the sense employed here of this relation, and this sense can then relate the senses expressed by the two descriptions without turning it into a thought that these senses are the same. Russell's argument against the meaning/denotation distinction as applied to descriptions only works insofar as it depends on a non-Fregean hybrid conception of a proposition which violates Frege's distinction between *Sinn* and *Bedeutung*.

The fact that Russell's argument against Frege is a failure does not, of course, vindicate Frege's treatment of descriptions as functional expressions or undermine Russell's own theory of descriptions. The key to this new theory is, as Russell put it later, that descriptions are "incomplete symbols," that is, phrases which "have no meaning in isolation" but are only "defined in certain contexts" (1910: 66). As such, for Russell, descriptions such as "the man" are to be regarded as essentially similar to the other denoting phrases discussed in "On Denoting," e.g. "a man" and "all men." They have "no meaning in isolation" in the sense that there is no thing (not even a concept) that is their "meaning" and that occurs as a constituent of the propositions expressed by sentences in which they occur. Instead they contribute to these propositions in more complex ways by fixing their structure, in the way that Russell conceives of the role of the universal quantifier as described above. It is therefore no surprise that on Russell's new theory of descriptions, the role of descriptions is elucidated by spelling out the quantificational structure of the propositions expressed by sentences in which they occur. This turns out to be a complex matter since Russell sticks to his initial assumption that all such propositions involve only the universal quantifier. But we can grasp the key points of Russell's new theory by providing just the first stage of it, which is that the proposition expressed by "The author of *Waverley* is Scott" is more clearly identified as the proposition expressed by the following sentence:

> For some x, (i) x is an author of *Waverley* and, (ii) for all y, if y is an author of *Waverley* then $y = x$, and, (iii) x = Scott.

Once the proposition is identified in this way solutions to Russell's two puzzles are immediate. There is no reason to identify the complex proposition thus expressed with the proposition that Scott is Scott, for the description "the author of *Waverley*" is not construed as giving rise to a complex name of a constituent within the proposition which turns out to be just Scott. So the distinction between significant and trivial identities is clearly preserved. Secondly, the role of empty descriptions is easily allowed for: the proposition expressed by "The present King of France is bald" is to be identified as that expressed by the sentence:

> For some x, (i) x is a present King of France and, (ii) for all y, if y is a present King of France then $y = x$, and, (iii) x is bald.

It is then unproblematic that there is such a proposition, and that it is "clearly false" as Russell declares that it should be.

The fact that these solutions are so straightforward, and do not depend on the underlying theory of propositions (though they are consistent with it), shows that one can abstract Russell's theory of descriptions from this underlying theory. This is in fact what has largely happened to Russell's theory of descriptions: it is taken to rest on the thesis that descriptions are quantifiers, and as such the Russellian position is best conceived of as one that employs a restricted "definite" quantifier to construe definite descriptions, so that the logical form of "the author of *Waverley* is Scott" can be better captured by construing it as

> For the x who is an author of *Waverley*, x = Scott.

One can then interpret Russell's reduction of the definite quantifier to other quantifiers as a misleadingly expressed way of spelling out the truth-conditions of this sentence. Once the matter is handled in this way, the debate with Frege can be re-opened, as a debate as to whether this way of construing descriptions is preferable to Frege's approach, according to which they are complex singular terms, comparable to functional expressions in mathematics, such that the logical form of "the author of *Waverley* was Scott" is captured by construing it as

> The author of (*Waverley*) = Scott.

This debate was famously revived by Strawson in his attack on Russell. Strawson introduced a range of linguistic data concerning the use of empty descriptions in situations which conflict with our "presupposition" that descriptions are non-empty, and argued on their basis that a Fregean position is in fact preferable to Russell's; in many cases, Strawson argued, we take it that the use of empty descriptions issues in statements that are "neither true nor false" and not "plainly false" in the way that Russell maintained. In fact, however, the linguistic data in this area are indeterminate, and most contemporary discussions focus instead on the relative merits of alternative accounts of the logical behavior of descriptions in complex sentences involving temporal modifiers and counterfactual constructions. Even when these are introduced, however, the issue remains surprisingly open – so much so that it seems to me best to conclude that definite descriptions blur the apparently sharp logical distinction between particular thoughts involving a singular term and general thoughts involving a quantifier.

Returning, however, to Russell himself I want finally to discuss the significance of the theory of descriptions within his philosophy generally. The first point is that Russell felt that his theory showed that there was no need to abandon his one-dimensional conception of meaning in favor of a Fregean theory with its all-encompassing distinction between *Sinn* and *Bedeutung*. Indeed, as we have seen, Russell felt that he could cope better than Frege with the "puzzles" of interesting identities and empty descriptions. This had an important implication for Russell's treatment of names: with no *Sinn*/*Bedeutung* distinction Russell was committed to the view that all identities involving just names are trivial and that there cannot be meaningful empty names. So all putative counterexamples had to be handled by supposing that the names in question were really just descriptions under disguise. The scope of this thesis was massively extended by a point that Russell introduces right at the end of "On Denoting" (p. 427), namely that our understanding of a proposition is based upon our "acquaintance" with its constituents, which are the meanings of the phrases used to express the proposition. For once one adds, as Russell does here, that we are not acquainted with matter and the minds of other people, it follows that our understanding of sentences that include putative names of material objects and other people cannot be achieved by identifying these things as the meanings of the names; instead we are bound to reinterpret the names as descriptions that invoke properties of things with which we are acquainted.

I shall discuss this radical doctrine of "knowledge by acquaintance" below. What I want to stress here is just that it was the theory of descriptions that made this doctrine tenable, since it seemed to offer a way of escaping the doctrine's otherwise unacceptable skeptical implications. The trick here was to suppose that "logical analysis" involving the theory of descriptions could save the appearances of common-sense belief even though its obvious foundations had been removed by the doctrine of limited acquaintance. This move became central to Russell's later "logical-analytic method in philosophy" (1914: v), to which I shall return below.

Another important point is the doctrine of "incomplete symbols," and in particular the central claim of the theory of descriptions that phrases whose meaning at first sight seems to consist in denoting some object turn out, after "logical analysis," not to have such a meaning at all. Instead their meaning is given only "in context": in the broader context of the sentences in which they occur. It is striking that Frege's context principle, which I mentioned earlier when discussing the issue of the unity of judgment, here enters into Russell's theory; but of course it does so only because descriptions are a counterexample to Russell's fundamentally non-contextual conception of meaning. For Russell the appeal to context is appropriate precisely where the demands of logic conflict with superficial grammar and its associated conception of meaning. The idea of such a conflict has been a deeply influential idea in the analytical tradition, giving rise to a very Platonist conception of "logical form" as something characteristically veiled by ordinary language. Wittgenstein rightly identified Russell's seminal role in developing this conception:

> All philosophy is a "critique of language" (though not in Mauthner's sense). It was Russell who performed the service of showing that the apparent logical form of a proposition need not be its real one. (*Tractatus Logico-Philosophicus* 4.0031)

Though by 1914 Russell took the view that "philosophy . . . becomes indistinguishable from logic" (in "On Scientific Method in Philosophy") he would have repudiated this characterization of philosophy as just "critique of language," since for him logic primarily concerns the logical forms of "the various types of facts" (*Papers*, 8: 65); indeed towards the end of his life, in the 1950s, he was very critical of the way in which philosophers seemed primarily concerned with language. As we have already seen, and shall see further below, Russell's logic was always shaped by metaphysical theses (e.g. concerning the nature of propositions) and driven by epistemological concerns (e.g. the doctrine of acquaintance). Nonetheless, his own theory of descriptions, and the uses to which he then put it in his logical-analytic program, did seem to many philosophers to show that it is through the logical analysis of language that philosophers can make progress in resolving old debates. To a considerable extent the whole project of analytical philosophy is founded upon this faith, and to that extent Russell's theory of descriptions remains, as Ramsey called it, "a paradigm of philosophy" (Ramsey 1931: 263n.).

Avoiding the contradiction

Russell's main concern in the years following the publication of *The Principles of Mathematics* was not in fact the theory of descriptions and its implications, which I have been discussing. Instead he was still preoccupied with his logicist project, on which he was now working with the Cambridge mathematician and (later) philosopher A. N. Whitehead, and he was, therefore, confronted by the need to avoid the perplexing contradiction he had discovered in 1901.

His first thought was that the conception of an incomplete symbol he had developed in connection with the theory of descriptions could be put to work to show what was wrong with the paradoxical set *R* of things that are not members of themselves. Russell developed this thought in an ingenious way by interpreting talk of sets in terms of the results of substitutions within propositions and then showed that, under this interpretation, the condition of self-membership cannot be coherently expressed. Regretfully, however, he decided that this approach was not the whole story since it did not resolve paradoxes concerning propositions, such as the liar paradox, which, he felt, were so closely related to his own paradox that there should be a single solution for them all.

He turned next to an idea that arose in the course of a debate with the French philosopher, Henri Poincaré, that these paradoxes arise only because the underlying argument tacitly involves a "vicious circle," in that something which has been defined in terms of a totality is then assumed to belong to this totality. Thus Russell's set *R* is defined in terms of the set of sets which are not members of themselves, and the contradiction is then arrived at by considering whether or not *R* belongs to this very set. Similarly in the case of the liar paradox the crucial move is that whereby the statement made by the liar is taken to be included in the scope of the liar's own statement. So, Russell thought, the way to avoid all these paradoxes is to adhere to the "vicious circle" principle, that "Whatever involves *all* of a collection must not be one of the collection" (1910: 37).

This principle gives rise to a hierarchy of "orders," since anything defined in terms of a collection of things of order n is held to be of order $n + 1$ and therefore not a candidate for membership of the first collection. In developing the idea further, however,

Russell went back to his earlier idea that the very definition of the paradoxical set *R* is somehow ill-formed. So he returned to the thought that our normal talk of sets involves incomplete symbols, and thus that sets are only "quasi-things" (1910: 81). But he no longer used his earlier interpretation of this talk in terms of substitutions within propositions: instead he introduced the conception of a "propositional function," a function whose values are propositions, and proposed a way of interpreting talk of sets in terms of propositional functions. Under this interpretation, talk of *a*'s membership of the set of things *x* such that *fx* is interpreted in terms of the truth of the proposition that is the value of the propositional function *f^x* (Russell's standard notation for propositional functions) for the argument *a* (i.e. the proposition *fa*). Hence the condition of self-membership entering into the definition of *R* is interpreted in terms of the truth of the proposition that is the value of a propositional function applied to itself as argument. But this, Russell claims, is incoherent: he takes it that the vicious circle principle implies that the propositions that are the values of a propositional function should in no case be specified by reference to the propositional function itself, and thus that

> there must be no such thing as the value for *f^x* with the argument *f^x* . . . That is to say, the symbol '*f(f^x)*' must not express a proposition. (1910: 40)

Russell reinforces this point by arguing that although there are functions of functions, in all cases functions must be of a different "type" from that of which they are functions; thus functions of simple individuals cannot themselves be individuals, but must have sufficient complexity to yield complete propositions when applied to individuals. Similarly functions of these functions, such as the quantifiers, have to have the type of complexity required to yield a complete proposition in these cases. Hence, he says, functions cannot be arguments to themselves, for they lack the type of complexity required to yield a complete proposition in this situation.

This line of thought generates a hierarchy of types (individuals, functions of individuals, functions of such functions, etc.) different from the hierarchy of orders generated by the vicious circle principle, which concerns the order of definitions. The resulting theory, the "ramified theory of types," is the result of merging these two hierarchies. There is an element of overkill in this theory, for there are now three reasons why the contradiction does not arise: (1) the vicious circle principle straightforwardly implies that *R* is of a higher order than its members and cannot therefore be a member of it; (2) Russell takes the principle to imply also that the definition underlying *R*, when spelled out in terms of propositional functions that apply to themselves, is ill-formed; (3) Russell also invokes a separate thesis that a function must have a different type of complexity from its arguments if it is to yield a complete proposition as value, which again implies that the definition underlying *R* is ill-formed.

The main aim of the theory, however, was not to avoid the contradiction but to fulfill the logicist project of providing a logical foundation for pure mathematics. This was in a way accomplished by Russell and Whitehead in their massive, though incomplete, trilogy *Principia Mathematica* (1910–13). They had found, however, that their task was obstructed by the complexities of the ramified theory. For example, at a crucial point in the standard theory of real numbers (the least upper bound theorem), the vicious circle principle is violated; hence Russell and Whitehead had, in effect, to set aside this principle by introducing the assumption (the "axiom of reducibility") that wherever a

propositional function is defined in terms of a totality to which it is then required to belong there is a way of defining it without reference to that totality, in terms of "predicative" functions that do not involve reference to the totality in question. As their critics observed, if Russell and Whitehead were going to help themselves to this assumption, then they could have simplified things a good deal by formulating the whole theory in terms of predicative functions in the first place. Indeed, after discussion with F. P. Ramsey Russell and Whitehead adopted a proposal of this kind in the second edition to *Principia Mathematica* (1927). The resulting theory, the "simple theory of types," no longer offers a solution to semantic paradoxes such as the liar paradox. But this is a positive gain since the vicious circle principle is anyway not a satisfactory resolution of these paradoxes, which depend on concepts such as truth whose complexities are separate from set theory itself.

But there are other problems, which afflict even this modified theory. Once individuals and functions (or sets) are divided into exclusive types, there has to be a separate, though isomorphic, arithmetic for each type, an idea that is highly counterintuitive. Furthermore the validity of standard arithmetic as applied to individuals requires the assumption that there is an infinity of such individuals. As Russell himself recognized (1919: 141), this assumption, or axiom, is not logical; it is clearly metaphysical (and may well be false).

To say this is to raise the question, central for the logicist project, as to what logic is. In *Principia Mathematica* logic is said to be the theory of formal inference, of inferences that depend merely on the logical form of the propositions involved. The distinction between "form" and "content" is then crucial; Russell takes it that the identification of the logical constants, by reference to which logical form is defined, is only a matter of enumeration. In his abandoned 1913 manuscript "Theory of Knowledge" he appreciates the need for some deeper theory, but remarks, "In the present chaotic state of knowledge concerning the primitive ideas of logic, it is impossible to pursue this topic further" (*Papers*, 7: 99). This is a surprising remark in the light of all Russell's work on logic. But there is no doubt that he had found the experience of coping with his contradiction a chastening experience, which had taught him that even in logic there are no simple answers, and thus that only "patience and modesty, here as in other sciences, will open the road to solid and durable progress" (*Papers*, 8: 73).

To compare logic with other sciences is to invite the question whether logic differs from them except in respect of its subject matter of formal inference, however exactly that be defined. This question is particularly apposite since at least until 1911 Russell affirmed that logic is synthetic, which might suggest that he also thought it is empirical. In fact, however, he held that it is a priori, and rests upon self-evident intuitions concerning the relationships between logical forms, which are universals (1912: ch. x). What is then a little odd is that he generally takes it that logical inference is just a matter of material implication (1910: 8–9), so that although logical inferences must be in fact truth-preserving it is not required that they preserve truth in all possible situations. His views in this area are not easily fitted together. Later, presumably under the influence of his former student Wittgenstein, he describes logical truths as "tautologies," which suggests a move to a conception of them as analytic; but in the "Lectures on the Philosophy of Logical Atomism" (published 1918) he is still very tentative:

> Everything that is a proposition of logic has got to be in some sense or other like a tautology. It has got to be something that has some peculiar quality, which I do not know how to define, that belongs to logical propositions and not to others. (*Papers*, 8: 211)

Modern set theories are based on the work of Zermelo and Von Neumann and avoid the problems that arise for Russell and Whitehead by sweeping away type distinctions. This implies that there is nothing ill-formed about the condition of self-membership, but Russell's paradox is avoided by denying that any well-formed condition (or propositional function, as Russell would call it) determines a set. Indeed it is standard to have an axiom of "foundation," which requires that the membership of a set be founded by being based upon "ur"-elements which do not themselves have members. This axiom (proposed by Mirimanoff in 1917) implies that the condition of non-self-membership does not determine a set and captures the intuition which lies behind Russell's "vicious-circle principle," but without obstructing the development of a systematic set theory that can be interpreted as a foundation for mathematics.

This does not, however, mean that modern set theory provides a vindication of Russell's logicist project. Russell's conception of a propositional function blurs the distinction between standard predicate logic and set theory, but, ironically, Russell's paradox itself shows the importance of maintaining a distinction here, and once this is drawn and set theory is provided with its own distinctive axioms concerning the existence of sets, there is no good reason to count set theory as logic. This does not mean that the logicist project is altogether untenable; for one can abandon set theory in favor of second-order logic and try to use this to provide a foundation for mathematics. Whether the resulting position is satisfactory remains disputed, but even if it is, it is still subject to the implications of Gödel's famous incompleteness theorem, which shows that arithmetic (and thus mathematics) cannot be completely captured within a formalized theory. If logic is just the theory of formal inference, as Russell maintained, then Gödel's theorem provides the ultimate refutation of his project of showing that "all mathematics is Symbolic Logic."

Logical atomism

When Russell was asked in 1924 to provide a personal statement of his philosophical position he chose to entitle it "Logical Atomism" (1956). This is a name that he began to use in 1914 (*Papers*, 8: 65) and then used in his 1918 "Lectures on the Philosophy of Logical Atomism." The rationale for the emphasis here on logic will be obvious; but what is the "atomism"? What are the "logical atoms"?

They are atomic facts. The reference here to "facts" is due to his abandonment by 1914 of his earlier Moorean conception of propositions. He has now adopted a form of the correspondence theory of truth according to which the truth of a proposition, now conceived of as normally a linguistic structure (though Russell also allows for imagistic mental propositions), is grounded in the perfect correspondence of logically simple propositions with atomic facts. I shall explain his reasons for this change of mind concerning propositions below, but since facts are said to be composed of the things that are the meanings of the words occurring in the proposition, it turns out that atomic facts differ little from old-style true elementary propositions, the propositions whose

truth formed the basis for the truth of propositions whose expression involves incomplete symbols (see "Lectures on the Philosophy of Logical Atomism," *Papers*, 8: 175).

Atomic facts are facts concerning the intrinsic qualities of, and relations between, particular individuals (*Papers*, 8: 177). But since the individuals and properties that constitute any fact we can talk about must be such that they are the meanings of the words we use, it follows that in practice the identity of atomic facts is constrained by the requirements of Russell's theory of meaning. And the crucial requirement here is that of our *acquaintance* with the things in question:

> A name, in the narrow logical sense of a word whose meaning is a particular, can only be applied to a particular with which the speaker is acquainted. (*Papers*, 8: 178)

Since Russell holds that we are not acquainted with ordinary things, such as Piccadilly and Socrates, but only with sensory particulars which are "apt to last for a very short time indeed" (*Papers*, 8: 181), it turns out that the atomic facts we can talk and think about do not deal with the familiar furniture of life, but for the most part only with the private objects of experience.

It is at this point that Russell's logical atomism makes contact with the doctrine of knowledge by acquaintance, which we encountered earlier, and thereby becomes a form of epistemological atomism. The core of that doctrine is expressed in his fundamental epistemological principle (in "Knowledge by Acquaintance and Knowledge by Description") that

> *every proposition which we can understand must be composed wholly of constituents with which we are acquainted.* (*Papers*, 6: 154)

With what things are we acquainted? Russell's approach is a combination of empiricism and rationalism. Through perception and introspection we gain acquaintance with particulars: with particular colors, noises, feelings, etc.; but through reflection we can also gain acquaintance with the universals of which these particulars are instances. This form of acquaintance involves "conception" (*Papers*, 6: 149–50), and is central to our capacity for intuitive awareness of a priori truths (1912: ch. X).

For our present purposes, the point on which to concentrate is our acquaintance with particulars. Russell uses Moore's general term "sense-data" for these sensory particulars but emphasizes that it is for him an open question whether such particulars may not also exist as unperceived "sensibilia." Sense-data so conceived are not "in the mind," though our awareness of them is immune from error. Initially Russell took the view that they are subjective because their relativity to their subject's position and condition implies that they cannot be combined in a single public world (1912: ch. I); but in 1914 he switched to the view that they are physical elements within private spaces, which are capable of being integrated into a single public world unless they belong merely to dreams and hallucinations. His account of this supposed integration is difficult and the details cannot be discussed here, but it is for him an important application of his "supreme maxim in scientific philosophising" (*Papers*, 8: 11) that

Wherever possible logical constructions are to be substituted for inferred entities.

The basic idea here is said by Russell to be similar to that involved in Ockham's razor, but in fact it is motivated by the need to escape the skeptical implications of his theory of knowledge by acquaintance. For where the apparent objects of putative knowledge are things of which we have no acquaintance, Russell's theory seems to imply that our beliefs can only ever be a matter of uncertain, speculative, inference concerning certain kinds of things "we know not what" whose existence is in some way implied by the existence of the things with which we are acquainted. Russell thinks he can avoid this skeptical result by substituting an alternative way of thinking, which can still be represented as a way of thinking about the apparent objects of knowledge but which, because it is "logically constructed" from thoughts of things (sense-data and universals) with which we do have acquaintance, does not require speculations concerning the existence of any further "inferred entities." This alternative strategy shows how knowledge is possible while respecting the fundamental epistemological principle quoted earlier.

Thus Russell takes it that knowledge of the external world is achieved on the basis of acquaintance with sense-data primarily by the application to these data of the logical principles involved in the theories of descriptions and of classes (though the integration of private spaces into a single public one involves more than logic). The result is that talk and thought putatively about public physical objects is to be provisionally interpreted in terms of logically complex propositions about classes of sense-data; but because classes are themselves "logical fictions" these propositions are themselves to be further interpreted in terms of propositional functions. Only then can we arrive at an indirect specification of the atomic facts, involving only items with which we are acquainted, of which we are aware, and which ground our knowledge of the external world. But, in principle, such a specification can be arrived at; so knowledge is possible, or so Russell supposes. In truth the matter is more tricky than he allows because of the "other minds" problem. As Russell acknowledges, the construction of the external world involves apparent reference to the sense-data of others, for the external world is essentially something that transcends our own sense-data. But since we are not acquainted with the sense-data of others, Russell's fundamental principle implies that we cannot understand any propositions requiring reference to them. So we have to start from our own sense-data alone and build out simultaneously to other minds and the external world. But whether this can be done in a way that meets Russell's requirements for knowledge is doubtful.

Russell's discussion of these matters connects directly with the logical positivist program of the 1920s and 1930s (especially Carnap's *Aufbau*), though, as we shall see below, he himself had modified his approach in important respects by then. That change was an indirect result of a different problem, which caused him to substantially rethink his position even while he was still developing his logical atomist program. The background to this is his abandonment of Moorean propositions. The reason for this change of mind (in 1906) is that Russell ceased to find it credible that there are "objective falsehoods," false propositions that are ontologically on a par with true propositions, that is, facts. As we have seen, he continued to accept the existence of facts, and

to that extent the old theory continued under a new name. But since propositions had been also taken to be the objects of judgment, false as well as true, Russell needed a new theory of judgment, which facts alone could not supply. His new theory, the "multiple-relation" theory, was that what had previously been conceived of as the constituents of a proposition that is the object of judgment should now be conceived of as terms of a new multiple-term relation conception of judgment; that is, instead of thinking of the Moorean proposition expressed by the sentence "Tom judges that A is larger than B" as having the logical form:

Judges (Tom, the proposition A is larger than B)

where this proposition is itself a complex entity somehow composed of A, B, and the relation of being larger than, we are to think of the same sentence (now regarded as a proposition itself, because it is to be the primary vehicle of truth and falsehood) as being such that, if true, it would correspond to a fact of the form:

Judges* (Tom, A, B, being larger than)

where "judges*" is the multiple-term relation that relates the subject of judgment (Tom) with certain objective terms (A, B, being larger than) in such a way that, together, they constitute a judgment that is true if and only if the objective terms constitute a fact – the fact that A is larger than B.

Russell never integrated this "no proposition" theory into his logical theory. Although it is stated in the introduction to *Principia Mathematica* (*PM*: 43–4), its implications for his conception of a propositional function, which has just been defined as a function whose values are propositions, are not worked through, nor are its implications for his substitutional treatment of quantifiers. Indeed it is flatly inconsistent with his theory of descriptions, since the complex interweaving of quantifiers and variables in that theory cannot be decomposed into "simple" constituents in the way required by the application of the multiple-relation theory to judgments involving descriptions.

Another difficulty comes from that old bugbear, the unity of judgment. For the theory in effect assumes that the objective terms of the multiple-term relation "judges*" can act as a surrogate for that which is judged, e.g. that A is larger than B. So the challenge that the terms by themselves do not constitute a complete judgment is one that cannot be avoided. Russell needs to explain how an appropriate specification of the right truth-making fact is determined simply by the objective terms of the multiple-term relation. One standard objection is that there is no basis for the distinction between judging that A is larger than B and that B is larger than A. It is arguable that this can in fact be handled simply by attending to the order in which the objects occur as terms of the relation "judges*"; but, as with descriptions, this strategy will not cope with general judgments involving multiple quantifiers – e.g. the judgment that all elephants are larger than all mice – where the bound variables obstruct the decomposition into simple constituents essential for Russell's approach. Wittgenstein's objection to Russell concerns a related point, that the theory permits one to judge nonsense. For unless some constraints are placed upon the terms of "judges*" there seems nothing to rule out a simple permutation of terms to generate, say,

Judges* (Tom, being larger than, A, B)

which would have to be a surrogate for Tom's "judgment" that being larger is A than B (see WITTGENSTEIN). Russell might seek to rule this out by placing type-restrictions on the terms of the multiple-term relation; but since type distinctions were explained in terms of the capacity of things to form a complete proposition Russell cannot appeal to them once he has embarked on his "no-proposition" perspective.

Russell was shattered by this objection. He had spent the months of May and June 1913 working at tremendous speed and in high spirits on a book about judgment and knowledge; but once he grasped Wittgenstein's point he abandoned the book (now published in *Papers*, 7) and fell almost into despair. Three years later he wrote to Ottoline Morrell about the crisis this event had induced:

> Do you remember that at the time when you were seeing Vittoz I wrote a lot of stuff about Theory of Knowledge, which Wittgenstein criticized with the greatest severity? His criticism, tho' I don't think you realized it at the time, was an event of first-rate importance in my life, and affected everything I have done since. I saw he was right, and I saw that I could not hope ever again to do fundamental work in philosophy. (1968: 57)

Later philosophy

Russell was prevented from lapsing into silence by the pressure of prior commitments at this time, most notably the Lowell lectures, which he delivered at Harvard in spring 1914 (published as *Our Knowledge of the External World*). Indeed his productivity during 1914 is a remarkable testament to his strength of will. Once the First World War began he turned with some relief from the need to rethink his philosophy to public opposition to the war, though by 1918 he was keen to return to philosophy. (At the very time in early 1918 that he was standing trial for his anti-war propaganda and then appealing against the terms of his sentence of six months' imprisonment he was also delivering the lectures on the philosophy of logical atomism (*Papers*, 8) I have referred to above).

Imprisonment, under the comfortable conditions permitted to him, turned out to provide the conditions of relative isolation that Russell needed to achieve a fresh start in his philosophy (though he also wrote his *Introduction to Mathematical Philosophy*, which is a lucid informal discussion of the main themes of *Principia Mathematica*). The starting point for this new work, which was published as *The Analysis of Mind* (1921), was William James's "neutral monism." Russell had already been thinking about this for some time; James's claim had been that the traditional opposition between mind and matter could be transcended by somehow conceiving of them as just different ways of thinking about something intrinsically neutral, which James called "experience." Russell, noting the similarities between this approach and his own account of the external world, develops a similar account based upon "sensations," which are like his old sense-data, except that he now holds that the fact that they are physical is no reason not to hold that they are not also mental (1995b: 143–4). The details of the constructions of the mind and of the physical world that follow are complex and unpersuasive; but what is nonetheless striking in the light of contemporary philosophy of mind is that

Russell sets out a position that combines ontological monism concerning mind and matter with an insistence that the natural laws regulating them are not reducible either way (1995b: 104–5), so that the position is not a reductive monism.

A central feature of Russell's analysis of mind is his attempt to do justice to what he takes to be the insights of scientific psychology, and in particular the behaviorist psychology being propounded at this time by John B. Watson. Russell cannot endorse the full behaviorist position; he thinks that beliefs typically involve mental imagery in a way that is incompatible with behaviorism. But he does endorse a broadly behaviorist conception of desire, and indeed refines it into a position that is recognizable as a precursor of contemporary functionalism. Furthermore he offers a causal account of the content of the images that enter into beliefs and extends this into a generally causal account of meaning. So his analysis of mind, including mental content, is based quite generally upon causal considerations, and this then provides him with the materials for a new theory of judgment to replace that which Wittgenstein had overthrown. He does not, however, take full advantage of this opportunity largely because he still thinks that the meaning of a complete sentence is constructed from the independent meanings of its constituent words (1995b: 273). So it was left to Ramsey to think the matter right through and propose, soon after, a tentative functionalist theory of judgment.

On the subject of knowledge, however, Russell clearly grasps the potential of his new causal conception of the mind. He begins *The Analysis of Mind* by rejecting his old conception of acquaintance, and later in the book he reinforces this break with his past by denying that we can obtain self-evident, certain, knowledge either by perception or by a priori intuition (pp. 262–6). In place of his old conception of knowledge, which, he now thinks, cannot rule out such "logically tenable, but uninteresting" skeptical hypotheses as that the world was created, with all our putative memories, five minutes ago (pp. 159–60), he offers "a more external and causal view" (p. 270) of knowledge. This is indeed precisely the view that is now familiar as "externalist"; and Russell introduces it by means of the now-familiar comparison between an accurate thermometer and someone with reliably true beliefs (p. 253ff).

In *The Analysis of Mind* Russell's presentation of this externalist conception of knowledge is somewhat tentative. In his last major work of philosophy, *Human Knowledge: Its Scope and Limits* (1948), Russell is much more assured and sophisticated in his development of this conception. His general aim here is Kantian: he seeks to explain how scientific knowledge is possible, but (unlike Kant) to do so within a broadly scientific conception of human life. The topic on which he then directs much of his attention is induction. This is a topic he had discussed in *The Problems of Philosophy*, where he had argued, first, that scientific knowledge is dependent upon the validity of the inductive principle that the greater the experience of the association of properties A and B the larger the probability that A and B will be found to be associated in new situations, and, secondly, that since this principle is presupposed in all reasoning from experience, it must be regarded as a self-evident a priori truth comparable to fundamental truths of logic. In his later work Russell begins by arguing that this position is untenable, because the inductive principle is open to obvious counterexamples if no restrictions are placed upon the properties involved. His argument here is similar to that later made famous by Nelson Goodman (as "The New Riddle of Induction") (see GOODMAN) and he takes from it the conclusion that this is not an area within which self-

evident truths are to be found. Instead, he proposes, it is essential to look to "scientific common sense" and discern the actual "postulates of scientific inference" (1948: 436).

In the last part of the book he undertakes this task in the context of a sophisticated conception of knowledge as something that comes in degrees: knowledge is not just true belief, but there are many kinds of warrant, some merely involving reliable connections, others involving understanding and reflection, which provide the higher degrees of knowledge. But in the end even the higher types depend on the existence of the connections that justify primitive types of knowledge. Russell's conclusion is that:

> Owing to the world being such as it is, certain occurrences are sometimes, in fact, evidence for certain others; and owing to animals being adapted to their environment, occurrences which are, in fact, evidence of others tend to arouse expectation of those others. By reflecting on this process and refining it, we arrive at the canons of inductive inference. These canons are valid if the world has certain characteristics which we all believe it to have. (1948: 514–15)

One could not ask for a clearer statement of the externalist's justification of induction, though Russell is under no illusion that this will altogether satisfy the philosophical skeptic.

This late work shows Russell still capable of originality at the age of 76. These late writings are often neglected today. But this is a mistake. For in these late writings, he practices the principle he had enunciated in 1924 that "we shall be wise to build our philosophy upon science" (1956: 339). As a result, his writings from this period connect directly with contemporary debates, since from the 1970s onwards the "naturalization" of analytical philosophy has introduced into philosophical debate the requirement of harmony with scientific knowledge that Russell had recognized fifty years earlier. Russell is still our contemporary.

Bibliography

Works by Russell

1903: *The Principles of Mathematics*, London: Allen & Unwin.
1910–13 (with A. N. Whitehead): *Principia Mathematica*, Cambridge: Cambridge University Press.
1912: *The Problems of Philosophy*, London: Williams & Norgate.
1914: *Our Knowledge of the External World*, London: Open Court.
1916: *Principles of Social Reconstruction*, London: Allen & Unwin.
1919: *Introduction to Mathematical Philosophy*, London: Allen & Unwin.
1948: *Human Knowledge: Its Scope and its Limits*, London: Allen & Unwin.
1956: "Logical Atomism," in *Logic and Knowledge*, ed. R. C. Marsh, London: Allen & Unwin, pp. 323–43. (First published in *Contemporary British Philosophy* (first series), ed. J. H. Muirhead, London: Allen & Unwin, 1924, pp. 359–83.)
1967: *The Autobiography of Bertrand Russell 1872–1914*, London: Allen & Unwin.
1968: *The Autobiography of Bertrand Russell 1914–1944*, London: Allen & Unwin.
1983–2000: *The Collected Papers of Bertrand Russell*, in 15 vols., London: Allen & Unwin/ Routledge. (The main papers cited in this chapter are listed below, and their page nos. in the *Papers* are used in text references.)

"Knowledge by Acquaintance and Knowledge by Description," in vol. 6, pp. 148–61. (First published in *Proceedings of the Aristotelian Society* 11 (1910–11), pp. 108–28.)
"Lectures on the Philosophy of Logical Atomism," in vol. 8, pp. 160–244. (First published in *The Monist* 28 (1918), pp. 495–527; 29 (1919), pp. 32–63, 190–222, 345–80.)
"On Denoting," in vol. 4, pp. 414–27. (First published in *Mind* 14 (1905), pp. 479-93.)
"On Scientific Method in Philosophy," in vol. 8, pp. 57–73. (Originally the Herbert Spencer Lecture, Oxford, November 1914.)
"Theory of Knowledge: The 1913 MSS," vol. 7.
"The Relation of Sense-data to Physics," in vol. 8, pp. 3–26. (First published in *Scientia* 16 (1914), pp. 1–27.)
1995a: *My Philosophical Development*, London: Routledge. (First published London: Allen & Unwin, 1955.)
1995b: *The Analysis of Mind*, London: Routledge. (First published London: Allen & Unwin, 1921.)

Works by other authors

Gödel, K. (1944) "Russell's Mathematical Logic," in *The Philosophy of Bertrand Russell*, ed. P. A. Schilpp, La Salle, IL: Open Court, pp. 123–54.
Griffin, N. (1991) *Russell's Idealist Apprenticeship*, Oxford: Clarendon Press.
Hylton, P. (1990) *Russell, Idealism, and the Emergence of Analytic Philosophy*, Oxford: Clarendon Press.
Monk, R. (1996) *Bertrand Russell*, vol. 1, London: Jonathan Cape.
Pears, D. F. (1967) *Bertrand Russell and the British Tradition in Philosophy*, London: Fontana.
Ramsey, F. P. (1931) *The Foundations of Mathematics and Other Logical Essays*, ed. R. B. Braithwaite, London: Routledge and Kegan Paul.
Watling, J. (1970) *Bertrand Russell*, Edinburgh: Oliver and Boyd.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

3

G. E. Moore (1873–1958)

ERNEST SOSA

Reflecting on his long philosophical career, G. E. Moore had this to say:

> I do not think that the world or the sciences would ever have suggested to me any philosophical problems. What has suggested philosophical problems to me is things which other philosophers have said about the world or the sciences.

Yet this philosophers' philosopher was lionized by the Bloomsbury literati, and his first book, *Principia Ethica*, is now included by the Modern Library Board among the one hundred most important nonfiction books of the century.

Born in 1873 to a middle-class family in a London suburb, Moore went up to Trinity College, Cambridge, at 19. After two years studying classics, he switched to philosophy under the influence of his friend Bertrand Russell, but soon after that it was Moore who led Russell from and against idealism. With "The Refutation of Idealism" Moore set the direction that was to take them both to logical and philosophical analysis, and to founding, along with Ludwig Wittgenstein and the logical positivists, the philosophical movement that came to be known as "analytic philosophy."

Moore's focus was not just on the giving of definitions or "analyses," however, though that certainly was central to his work, as it was for Plato. On the contrary, he made it clear that it is also a job for philosophy to give "a general description of the *whole* of this universe, mentioning all the most important kinds of things which we *know* to be in it." He also reflected long and deeply on what knowledge is and on how it might be attained. These reflections had an important impact on Wittgenstein, whose *On Certainty* is a response to Moore in epistemology (see WITTGENSTEIN). And it was ethics that first attracted Moore's intensely concentrated, patient attention, yielding that great first book of his.

In that book, Moore introduces the analysis of concepts and properties into components. Consider, for example, being male and being a sibling, which come together through conjunction to form the complex concept of being a brother. According to Moore, some concepts are not thus analyzable, however, among them that of being yellow and that of being good (i.e. intrinsically good), the latter of which is fundamental to ethics.

This fundamental concept of ethics is said to be not only simple but also nonnatural, and irreducible to any natural property. Thus it is not analyzable in utilitarian style as

a matter of causing or containing states of pleasure or pain. There are ways of being intrinsically good not dreamt of by such hedonism; being good cannot be identified with any one member of the plurality of intrinsic goods. Moreover, concerning any natural property *X* proposed as identical with the property of being good, there is always some such "open question" as: "Is having *X* inevitably necessary and sufficient for being good?" Even if we answer such a question in the affirmative, once our answer is disputed, as surely it will be if not absolutely trivial, that will show the property *X* and the property of being good *not* to be one and the same property. For if these *were* the same, then our question would be tantamount to the question "Is having *X* inevitably necessary and sufficient for having *X*?" – than which few questions can be more trivial. Surely this question cannot occasion the sort of controversy that always attends theories of the good.

Moore puts forward some excellent issues here, but his answers are not entirely satisfactory if only because based on too few distinctions (uncharacteristically so). At a minimum, we need the distinction between meaning analysis and philosophical, metaphysical analysis, as well as the distinction between concepts and properties. Perhaps "male sibling" can help provide a meaning analysis of "brother," so that the question "Is a brother a male sibling?" is indeed trivial. By contrast, "is productive of a greater balance of pleasure over pain than any alternative" may not give us a *meaning* analysis of "is the right thing to do." But it is left open that it may give a good philosophical analysis in any case, if one that is *not* obvious, not just a matter of surface meaning. How then are we to think of philosophical analysis compatibly with the fact that, unlike a kind of (surface) meaning analysis, it is far from trivial and requires reflection? This has proved a difficult and troubling problem for analytic philosophy, and has been not so much solved as shelved under the heading of "the paradox of analysis." Those who still care about piecemeal analysis – which still includes many philosophers, though not all who count themselves "analytic" – have good reason to feel nagged by this worry.

Not through such cool analytic work did Moore attract the Bloomsburies. Most likely it was rather through bucking the Victorian penchant for rigid rules. These could not withstand Moore's probing intelligence except as rules of thumb. But Moore also distanced himself from Bentham–Mill utilitarianism. It is certainly not just pleasure that fundamentally deserves our admiration and pursuit. There are things of various sorts that separately have that special status of the *intrinsically* good; prominent among these figure, first, the enjoyment of certain human relations and, second, the appreciation of things of beauty. Here Moore's milieu and upbringing may show. Talk to the starving millions about things of beauty, and you will be less successful than in Bloomsbury. Still, pluralist that he is, Moore could simply accept further intrinsic goods that his list may have overlooked, without any major setback to his overall position.

An act is *right*, Moore advises, if and only if there is no better available alternative, where the value of each alternative is a function not only of its own intrinsic merit but also of the combined value of its consequences unto eternity. These combinations are not just brute additions, however, but may involve special value deriving from the way in which they combine, as when someone is rewarded for doing well or good, or when someone is unrewarded or punished for doing ill or harm.

Especially in his metaphysics and epistemology, Moore joined a tradition of common sense philosophy, one to which he was no doubt inherently and antecedently receptive.

Perhaps he came to know it so deeply during his years in Edinburgh, between the end of his fellowship at Trinity in 1904 and the beginning of his tenure back in Cambridge as a lecturer in 1911. The fuller name of that tradition is, after all, Scottish Common Sense, which is explained mostly by the fact that its greatest early proponent was the Scotsman Thomas Reid. In any case, however he may have been led to this tradition, Moore took to it naturally, and would defend it and develop it in his own inimitable style.

For Moore serious analysis required a kind of patient illumination of detail and nuance which is very difficult to follow attentively, and impossible for nearly anyone but Moore himself to produce. Indeed many of his essays in metaphysics and epistemology are made up almost entirely of such minute analysis, leading up to a few brilliant insights very quickly stated near the end. Take, for example, his famous "Proof of an External World." Except for the concluding few pages, Moore is engaged mostly in figuring out in great detail what "externality" could mean. That he is able to stay with that question through so many twists and turns, and that he does not bother to hide the analytic complexities illuminated by his intelligence, shows his integrity, but also makes him hard to read. If one stays until the end, however, the reward comes in the insights of the concluding pages. This is true not only of "Proof of an External World," but also of "Four Forms of Scepticism," wherein he is again defending his common sense from skeptical attack.

That commonsensical view had already been expounded in his "Defence of Common Sense," wherein he describes various features of the world as he commonsensically believes it to be. At the center of these ontological reflections Moore inquired into the nature of sensory experience and its relation to physical reality in a way characteristically exploratory and attentive to detail. Are sense-data identical with physical surfaces? Are they rather nonphysical denizens of our mental world while representative of physical realities beyond them? Is physical reality itself to be viewed as somehow a construction from or analyzable or reducible to combinations of sense-data? Moore long struggled with such questions, but his work in this area never reached closure.

Nevertheless, he felt certain enough of the core of his common sense, whatever its correct analysis may turn out to be, that he was willing to give it firm expression through a list of some of its central commitments, among them the following two: that he has and has for some time had a human body, which has been in contact with the surface of the earth, and that there have been many other three-dimensional things at various distances from his body. Such propositions form a first group. In a second group are such propositions as that he has had experiences of various sorts, and that he has observed various things in his surroundings at the time, and has had dreams, and other mental states. Finally, in a third group is the proposition that the same is true of many other human beings. In that paper Moore also quite explicitly claims, finally, that he *knows with certainty* propositions in his first two groups, and that the many other human beings of whom similar things are true also have frequently enjoyed such certain knowledge.

These are of course the claims that set up his confrontation with the skeptic, where Moore's legendary patience and powers of analysis are very much in evidence. It is this work, in my judgment, that manifests a depth of insight beyond anything shown in

Moore's other work. Not that we cannot see gaps and problems in hindsight. But his achievement was nonetheless real and impressive, and impressed the singularly unimpressionable Wittgenstein, from whom it elicited his own best work in epistemology. In the next and final part of this discussion, we turn to this work of Moore's.

Is the existence of external things just an article of faith? Certainly not, says Moore, and offers us a proof (which is here simplified), thus aiming to remove Kant's "scandal to philosophy."

Moore's proof Here is a hand (a real, flesh and bone hand).
Therefore, there is at least one external thing in existence.

According to Moore, his argument meets three conditions for being a proof: first, the premise is different from the conclusion; second, he knows the premise to be the case; and, third, the conclusion follows deductively ("Proof of an External World," in 1962: 144–5). Further conditions may be required, but he evidently thinks his proof would satisfy these as well.

As Moore is well aware, many philosophers will feel he has not given "any satisfactory proof of the point in question" (1962: 147). Some, he believes, will want the premise itself proved. But he has not tried to prove it, and does not believe it can be proved. Proving that here is a hand requires proving one is awake, and this cannot be done.

Does Moore adequately answer the skeptic? Many have denied it for the reason that he fails to rule out a crucial possibility: that our faculties are leading us astray, for example that we are dreaming. Aware of this objection, Moore grants, in "Certainty," that to know he is standing he must know he is awake ("Certainty," in 1962). The point "cuts both ways," however, and he would prefer to conclude that he *does* know he is awake since he *does* know he is standing.

This has persuaded nearly no one. On the contrary, some have thought him committed to an argument, M below, which is like Argument A, preceding it:

Argument A
A1 This map is a good guide to this desert.
A2 According to the map an oasis lies ahead.
A3 Therefore, an oasis lies ahead.

Argument M
M1 My present experience is a veridical guide to reality (and I am not dreaming).
M2 My present experience is as if I have a hand before me.
M3 Therefore, here (before me) is a hand.

When challenged on premise A1, our desert dullard responds: "I must know A1 since the only way I could know A3 is through argument A, and I *do* know A3." Is this a just comparison? Is Moore's response to the skeptic relevantly similar?

If Moore depends on argument M for his knowledge of M3, his response seems like the dullard's. The dullard is wrong to respond as he does. He must say how he knows his premise without presupposing that he already knows the conclusion. And Moore would seem comparably wrong in the analogous response to the skeptic. In explaining how he knows M1, he must not presuppose that he already knows M3.

Does Moore depend on argument M for his knowledge of M3? There is reason to think that he does not, given his emphatic acknowledgment that he cannot *prove* M3. After all, M would seem a proof of M3 just as good as Moore's own "proof of an external world." Moore concedes, in effect, that *if* he does not know that he is not dreaming *then* he does not know of the hand before him. But that is *not* necessarily because he takes himself to know M3 only through M or any other such argument. And, in any case, even if he is relying on some such argument, which would require making that concession, the defender of common sense has other options.

One might, after all, make that concession only because of the following "principle of exclusion":

PE If one is to know that *h*, then one must exclude (rule out) every possibility that one knows to be incompatible with one's knowing that *h*.

As Moore grants explicitly, the possibility that he is just dreaming is incompatible with his knowing (perceptually) that he has a hand before him. And this, in combination with PE, is quite sufficient to explain his concession above.

Suppose Moore is not depending on argument M for his knowledge of M3. Although he recognizes his need to know he is not dreaming, suppose that is only because he accepts PE, our principle of exclusion. Then the sort of ridicule cast on the dullard is misdirected against Moore. What is more, it is not even clear that Moore must know *how* he knows he is not dreaming if he is to know M3. That is not entailed by the application of the principle of exclusion. All that follows from the application of that principle is that Moore must know *that* he is not dreaming, not that he must know *how* he knows this.

In fact, however, the historical Moore *did* rely on something very much like argument M (more on this below). So is he not after all exposed to the damaging comparison with the desert dullard?

Not at all. There seems no good reason why, in responding to the skeptic, Moore must *show how* he knows he is not dreaming. Of course his response to the skeptic would be *enhanced* if he *could* show that. But it now seems *not* properly subject to ridicule even if he is not then in a position to show how he knows he is not dreaming. The question he is addressing is *whether* he knows that he is not dreaming, and, at most, by extension, what grounds he might have for his answer to *that* question, in answering which he does not, nor need he, *also* answer the question of *how* he knows himself to be awake and not dreaming.

It might be replied that one cannot know that here is a hand if one's belief rests on the unproved assumption that one is awake. According to Moore, however, things which cannot be proved might still be known. Besides, even though he cannot prove that he is awake, he has "conclusive evidence" for it. Unfortunately he cannot state his evidence, and the matter is left in this unsatisfactory state at the end of "Proof of an External World." But Moore has more to say in another paper of the period, "Four Forms of Scepticism" (1962: 193–223). There he takes himself to know for sure about the hand before him, and takes this knowledge to be based on an inductive or analogical argument. We are told that introspective knowledge of one's own sensory experience can be immediate, unlike perceptual knowledge of one's physical surroundings. While agreeing with Russell that one *cannot* know *immediately*

that one sees a hand, Moore thinks, *contra* Russell, that he *can* know it *for certain*. And he disagrees with Russell more specifically in allowing knowledge for certain about his hand through analogical or inductive reasoning from premises known introspectively.

However, it is doubtful that any allowable form of inference – whether deductive, inductive, or analogical – will take us from the character of our experience to the sort of knowledge of our surroundings that we ordinarily claim.

Familiar skeptical scenarios – dreaming, evil demon, brain in a vat, etc. – show that our experience prompts but does not logically entail its corresponding perceptual beliefs. Experience as if there is a fire before us does not entail that there is a fire there, experience as if here is a hand does not entail that here is a hand, etc. Perhaps what is required for one's beliefs and experiences to have certain contents entails that these could not possibly be *entirely* false or misleading. Indeed, some such conclusion follows from certain externalist and epistemic requirements on one's justified attribution of familiar contents to one's own experiences or beliefs. But even if that much is right – which is still controversial – one's experience or belief that here is a hand, or yonder a fire, might still be wildly off the mark. We cannot deduce much of our supposed knowledge of the external from unaided premises about our experience.

As for inductive or analogical reasoning, only abductive reasoning – inference to the best explanation – offers much promise, but it seems questionable as a solution to our problem.[1] Suppose (1) that we restrict ourselves to data just about the qualitative character of our own sensory experience, and (2) that we view belief in a commonsensical external world as a theory postulated to explain the course of our experience. What exactly is the proposal? Is it proposed that when ordinarily we accept the presence of a hand before us, we *do* know, and know on the basis of an abductive inference; or is it proposed rather that in such circumstances we have resources that *would* enable us to know if only we used those resources to make effective abductive arguments? The second, more modest, proposal is *too* modest, since it leaves our ordinary perceptual beliefs in a position like that of a theorem accepted through a guess or a blunder, one that we do have the resources to prove after much hard thought, but one that we have not come close to proving at the time when we are just guessing or blundering.

Even the modest proposal, moreover, seems unlikely to succeed. *Could* we form a rich enough set of beliefs purely about the qualitative character of our sensory experience, one rich enough to permit abductive inferences yielding our commonsense view of external reality? This seems doubtful when we consider (1) that such pure data beliefs could not already presuppose the external reality to be inferred, and (2) that the postulated commonsense "theory" of external reality must presumably meet constraints on abductive inference: for example that the postulated theory be empirically testable and also simpler and less *ad hoc* than alternatives (e.g. Berkeley's). These requirements plausibly imply that our data must go beyond detached observations, and include some acceptable correlations. Yet these correlations are unavailable if we restrict ourselves to beliefs about the character of our experience.[2] Most especially are they unavailable, and most especially is the postulated inference implausible, when our database is restricted, as it is by Moore, to introspectively known facts of one's own *then present* subjective experience, and to *directly recalled* facts of one's own earlier experience. (If

deprived of the epistemic resources of testimony and of retentive memory – except insofar as such resources can be validated by reason-cum-introspection, which is not very far if at all – then there is precious little we can any longer see ourselves as knowing, thus deprived.)

Accordingly, the skeptic has a powerful case against Moore's claim that our knowledge of the external is based on an inductive or analogical inference from such information about our experience. It is not realistic to suppose that we consciously make such inferences in everyday life. It is more plausible to conceive of such inferences as implicit or dispositional, but even this strains belief. Besides, even granted that we make such inferences if only implicitly, do they yield simpler and less *ad hoc* hypotheses than alternatives? That is far from clear; nor do such hypotheses seem empirically testable and credible simply as explanations of the purely qualitative character of our then present or directly recalled experience.

Having reached a dead end, let us have some second thoughts on Moore's view of perceptual beliefs as inferential. Here he joined a venerable tradition along with Russell himself. If perceptual knowledge is thus mediate and inferential, what knowledge can qualify as immediate and foundational? Modern philosophy begins with Descartes's canonical answer to this question.[3]

Descartes had two circles, not only the big famous one involving God as guarantor of our faculties, but also a smaller one found in the second paragraph of his Meditation III, where he reasons like this:

> I am certain that I am a thinking being. Do I not therefore also know what is required for my being certain about anything? In this first item of knowledge there is simply a clear and distinct perception of what I am asserting; this would not be enough to make me certain of the truth of the matter if it could ever turn out that something which I perceived with such clarity and distinctness was false. So I now seem to be able to lay it down as a general rule that whatever I perceive very clearly and distinctly is true.[4]

About the *cogito,* I wish to highlight the inference drawn by Descartes: *So I now seem to be able to lay it down as a general rule that whatever I perceive very clearly and distinctly is true*. Just what is Descartes's argument in support of this general rule? Would his reasoning take the following form?

1 Datum: I know with a high degree of certainty that I think.
2 I clearly and distinctly perceive that I think, and that is the only, or anyhow the best account of the source of my knowledge that I think.
3 So my clear and distinct perception that I think is what explains why or how it is that I know I think.
4 But my clear and distinct perception could not serve as a source of that knowledge if it were not an infallibly reliable faculty.
5 So, finally, my clear and distinct perception must be an infallibly reliable faculty.

The move from (1) and (2) to (3) is an inference to an explanatory account that one might accept for the coherence it gives to one's view of things in the domain involved. Descartes does elsewhere appeal to coherence at important junctures.[5] So he may

be doing so here as well, although questions do arise about how Descartes views coherence. Does he accept the power of coherence to add justified certainty, and, in particular, would he claim infallibility for (sufficiently comprehensive and binding) coherence as he does for clear and distinct intuition?[6] In any case, the comprehensive coherence of his world-view would be enhanced by an explanation of how clear and distinct perception comes to be so highly reliable, even infallible. And this is just what Descartes attempts, through his theological and other reasoning. Descartes can see that reason might take him to a position that is sufficiently comprehensive and interlocking – and thereby defensible against any foreseeable attack, no holds barred, against any specific doubt actually pressed or in the offing, no matter how slight. Unaided reason might take him to that position. Need he go any further? What is more: might one reach a similar position while dispensing with the trappings of Cartesian theology and even of Cartesian rationalism?

Compare now how Moore might have proceeded:

1 Datum: I know with a high degree of certainty that here is a hand.
2 I can see and feel that here is a hand, and that is the only, or anyhow the best account of the source of my knowledge that here is a hand.
3 So my perception that here is a hand is what explains why or how it is that I know (with certainty) that here is a hand.
4 But my perception could not serve as a source of that degree of justified certainty if it were not a reliable faculty.[7]
5 So, finally, my perception must be a reliable faculty.

Moore could of course go on to say more about the nature of the perception that assures him about the hand. He might still say that such perception involves an implicit inference from what is known immediately and introspectively, perhaps an inductive or analogical inference of some sort. And that might make his view more comprehensively coherent, but we have already seen reasons why postulating such an inference is questionable. So we focus rather on a second alternative: Moore might well take perceiving to involve no inference at all, not even implicit inference, but only transfer of light, nerve impulses, etc., in such a way that the character of one's surroundings has a distinctive impact on oneself and occasions corresponding and reliable beliefs. This might also amount eventually to a comprehensively coherent view of one's knowledge of the external world. *And its epistemologically significant features would not distinguish it in any fundamental respect from the procedure followed by Descartes.*

There are other ways of opposing Moore besides that of the traditional skeptic. These are all based in some way or other on a key requirement of "sensitivity" for knowledge, one imposed on any belief candidate for knowledge, as follows: one's belief that *p* amounts to knowledge that *p* only if one would *not* believe that *p* if it were not the case that *p*.

It is initially very tempting to accept the sensitivity idea common to the various forms of sensitivity-based opposition to Moore: namely, the skeptical, tracking, relevant-alternative, and contextualist approaches that share some form of commitment to that requirement. And, given this idea, one can then argue powerfully for the first premise of the skeptic's "argument from ignorance," AI, formulable by means of the following abbreviations:

H I am a handless brain in a vat being fed experiences as if I were normally embodied and situated (see PUTNAM).
G I now have hands.

Here now is AI:

(i) I do not know that not-H.
(ii) If (i), then C (below).
C I do not know that G.

That lays out the skeptic's stance. Moore for his part grants the skeptic premise (ii), but rejects C and therefore (i). Robert Nozick's stance is different.[8] Like Moore, he rejects C. Like the skeptic, he affirms (i). So he must reject (ii), which he does aided by his independently supported account of knowledge as tracking. It is not only Nozick who rejects closure under known entailment; so does the relevantist, for whom in order to know some fact X you need not know (and often cannot know) the negation of an alternative known to be incompatible with X, so long as it is not a "relevant" alternative.

Having granted to the skeptic his premise A(i), contextualism is able to defend ordinary claims to know only by distinguishing the ordinary contexts in which such claims are made from the context where the skeptic asserts his distinctive premise in the course of giving argument AI. With this difference in context comes a difference in standards; and, because of this difference, it is incorrect to say in the skeptic's context that one knows G, correct though it may be to say it in an ordinary context.

That response to the skeptic faces a problem. Moore's opponents argue that sensitivity is necessary for correct attributions of knowledge.[9] Despite its plausibility, however, serious objections have been published against any such requirement of sensitivity. But the problems for sensitivity do not affect a similar requirement of "safety." A belief is sensitive iff had it been false, S would not have held it (i.e. it would have been false only without S holding it), whereas a belief is *safe* iff S would have held it only with it being true. For short: S's belief B(p) is sensitive iff $\sim p \rightarrow \sim B(p)$, whereas S's belief is safe iff $B(p) \rightarrow p$. These are not necessarily equivalent, since subjunctive conditionals do not contrapose.[10]

And now we see the problem faced by the contextualist response to the skeptic: namely, that an alternative explanation is equally adequate for undisputed cases (undisputed, for example, between those who opt for a Moorean stance opposing the skeptic's distinctive premise (i) and those who opt for a contextualist stance which accepts it). According to this alternative explanation, it is safety that (correct attribution of) "knowledge" requires, a requirement violated in the ordinary cases cited, wherein the subject fails to know. One fails to know in those cases, it is now said, because one's belief is not safe. Suppose this generalizes to all uncontentious cases adduced by the contextualist to favor his sensitivity requirement. Suppose in all such cases the condition required could just as well be safety as sensitivity. And suppose, moreover, that the problems for sensitivity briefly noted above do not affect safety, as I have claimed. If so, then one cannot differentially support sensitivity as the right requirement, so as to invoke it in support of the skeptic's main premise.

Here is the striking result: if we opt for safety as the right requirement then a Moorean stance is defensible, and we avoid skepticism.[11] That is to say, one does satisfy

the requirement that one's belief of not-H be safe: after all, one *would* believe that not-H (that one was not so radically deceived) only if it was true (which is not to say that one *could* believe that not-H only if it was true). In the actual world, and for quite a distance away from the actual world, up to quite remote possible worlds, our belief that we are not radically deceived matches the fact whether or not we are radically deceived.[12]

Consider, moreover, the need to explain how the skeptic's premise – that one does not know oneself not to be radically misled, etc. – is as plausible as it is. That requirement must be balanced by an equally relevant and stringent requirement: namely, that one explain how that premise is as *implausible* as it is.[13] To many of us it just does not seem so uniformly plausible that one cannot ever be said correctly to know that one is not then being fed experiences while envatted. So the explanatory requirement is in fact rather more complex than might seem at first. And, given the distribution of intuitions here, the contextualist and the Nozickian still owe us an explanation.

Interestingly, our distinction between sensitivity and safety may help us meet the more complex explanatory demand, compatibly with the Moorean stance, which I adopt as my own. My preferred explanation may be sketched as follows.

Those who find the skeptic's distinctive premise plausible *on the basis of the sorts of sensitivity considerations favored by opponents of Moorean common sense* may perhaps be confusing sensitivity with safety, and may on that basis assess as correct affirmations of that premise. After all, the requirement of safety is well supported by the sorts of considerations adduced by Moore's opponents. Sensitivity being so similar to safety, so easy to confuse, it is no surprise that one would find sensitivity so plausible, enough to mislead one into assessing as correct affirmations of that premise.

The plausibility of the skeptic's premise is thus explained compatibly with its falsity, which fits the stance of the Moorean. Of course all we really need in order to explain the plausibility of the skeptic's premise is that it clearly enough follows from something plausible enough. And the sensitivity requirement may perhaps play that role well enough independently of whether it is confused with a safety requirement. But that would still leave the question of why sensitivity is so plausible if it is just false. And here there might still be a role for safety to play: if this requirement of safety is plausible because it is true and defensible through reflection, then it may be deeply plausible to us simply through our ability to discern the true from the false in such a priori matters. Compatibly with that, some of us may be misled into accepting the requirement of sensitivity because it is so easy to confuse with the correct requirement, that of safety.[14]

I have wanted to convey the power and depth of Moore's thought not only by describing it at a lofty distance but also by engaging with it at close quarters. Despite the reservations I have recorded on this or that point, I hope to have made it clear how persuasively right are his views on some of the most difficult and disputed issues in the history of our subject. But being right does not alone confer greatness in philosophy. From his earliest days as a thinker, Moore was not only right but also able to think for himself in ways opposed to the regnant orthodoxies, and to prevail as a master dialectician. One main source of his influence is now impossible to capture fully, however, since it resided in his *viva voce* contributions to the intellectual life of that golden age of

Cambridge philosophy. This Socratic side is well conveyed by his younger Cambridge colleague, C. D. Broad, in an obituary for Moore:

> It was by his lectures, his discussion-classes, his constant and illuminating contributions to discussion at the Cambridge Moral Science Club and the Aristotelian Society, and his private conversation with his colleagues and pupils that he mainly produced his effects on the thought of his time.

Notes

1 For Russell the "common sense hypothesis" of independent physical objects is "simpler" than the supposition that life is but a dream (as he explains in chapter II of *The Problems of Philosophy*). For Quine the "hypothesis of ordinary physical objects" is "posited" or "projected" from the data provided by sensory stimulations. "Subtracting his cues from his world view, we get man's net contribution as the difference" (*Word and Object* (Cambridge, MA: MIT Press, 1960), p. 5). That Quine's position is deeply problematic is shown by Stroud (*The Significance of Philosophical Skepticism* (Oxford: Oxford University Press, 1984), ch. VI).

2 This is argued by Wilfrid Sellars in "Phenomenalism," in his *Science, Perception, and Reality* (London: Routledge and Kegan Paul, 1963) (see SELLARS).

3 The shift to discussion of Descartes may seem abrupt; however, what we find about the nature of immediate knowledge in that discussion has important implications for a position that Moore failed to explore. Skeptics who are willing to grant Descartes his immediate knowledge through introspection or rational intuition would need to explain exactly why perception could never yield such knowledge. (And what of memory?) The discussion of Descartes to follow is meant to highlight this issue.

4 *The Philosophical Writings of Descartes*, ed. J. Cottingham, R. Stoothoff, and D. Murdoch (Cambridge: Cambridge University Press, 1975), vol. II, p. 24.

5 In his *Principles of Philosophy* (Part IV, art. 205) for example, he notes that if we can interpret a long stretch of otherwise undecipherable writing by supposing that it is written in "one-off natural language," where the alphabet has all been switched forward by one letter, etc., then this is good reason for that interpretation. There he also argues for his scientific account of reality in terms of certain principles by claiming that "it would hardly have been possible for so many items to fall into a coherent pattern if the original principles had been false" (Cottingham et al., *Philosophical Writings of Descartes*, p. 290). Of course, if we join Descartes in adopting this sort of inference to an account that aids comprehensive coherence we will need to be able to distinguish it relevantly from the rejected abductive inference to an external world from introspective data about one's own experience and direct memories about one's past experiences. But there are important differences: for one thing, the present Cartesian inference is not an inference to a causal account, one with discernible rivals that we are unable to rule out without vicious circularity. But it remains to be seen whether the *additional* theological project that Descartes next launches is or is not open to similar problems as those that beset the abductive inference to the external world, or even worse problems. We do not consider these issues which are matters of detail by comparison with the more abstract epistemological structure of Descartes's reasoning that we consider.

6 My attribution to Descartes is tentative because of the enormous bibliography on the "Cartesian circle." In deference to that important tradition of scholarship, I do no more than *suggest* that there is logical space for an interpretation of Descartes that is perhaps more

complex than many already tried, but that seems coherent and interesting. (I am myself convinced that this *is* Descartes's actual position, and defend this more fully elsewhere.)

7 Here one would reduce Descartes's requirement of *infallible* certainty.

8 Robert Nozick, *Philosophical Explanations* (Cambridge, MA: Harvard University Press, 1981).

9 Keith DeRose, "Solving the Skeptical Problem," *Philosophical Review* 104 (1995), pp. 1–52.

10 If water now flowed from your kitchen faucet, it would *not* then be the case that water so flowed while your main valve was closed. But the contrapositive of this true conditional is clearly false.

11 I mean that *we* in our reflection and in our discussions in journal and seminar, avoid skepticism; we can say right here and now that we do know various things, and not just that we say "I know" correctly in various contexts not now our own.

12 This sort of externalist move has been widely regarded as unacceptably circular, mistakenly, as I argue in detail elsewhere.

13 When I have asked my classes to vote on that premise, generally I have found that those who find it false outnumber those who find it true, and quite a few prefer to suspend judgment. At every stage people spread out in some such pattern of three-way agreement-failure.

14 For a fuller defense of a Moorean stance in epistemology by comparison with alternative ideas on the epistemology marketplace, see my "How to Defeat Opposition to Moore," *Philosophical Perspectives 13; Epistemology* 13 (1999), Supplement to *Noûs*, pp. 141–55.

Bibliography

Works by Moore

1903: *Principia Ethica*, Cambridge: Cambridge University Press.
1912: *Ethics*, London: Williams & Norgate.
1922: *Philosophical Studies*, London: Routledge and Kegan Paul.
1953: *Some Main Problems of Philosophy*, London: Allen & Unwin.
1962: *Philosophical Papers*, New York: Collier Books.

Works by other authors

Baldwin, T. (1990) *G. E. Moore*, London and New York: Routledge.
Fratantaro, S. (1998) *The Methodology of G. E. Moore*, Aldershot and Brookfield, VT: Ashgate Publishing Ltd.
Klemke, E. D. (1969) *The Epistemology of G. E. Moore*, Evanston, IL: Northwestern University Press.
Schilpp, P. A. (1968) *The Philosophy of G. E. Moore*, La Salle, IL: Open Court.
Stroll, A. (1994) *Moore and Wittgenstein on Certainty*, Oxford: Oxford University Press.
White, A. R. (1958) *G. E. Moore*, Oxford: Blackwell Publishers.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

4

C. D. Broad (1887–1971)

JAMES VAN CLEVE

Charlie Dunbar Broad was a leading contributor to analytic philosophy of the twentieth century, known not so much for any startlingly original doctrines he propounded as for his formidable powers of distinction, analysis, and argument. Born in London, he was educated at Dulwich College and Cambridge. He entered Cambridge in 1905, first studying physics and chemistry in the natural science tripos and then switching to philosophy in the moral science tripos. The influence of Russell and Moore at Cambridge was then very strong and shows itself in Broad's work (see RUSSELL and MOORE). He published his dissertation as *Perception, Physics, and Reality* in 1914. For a period of years beginning in 1911 he served as G. F. Stout's assistant in St. Andrews, and in 1920 he was appointed professor at the University of Bristol, where he gave the course of lectures in philosophy for natural science students that became *Scientific Thought*. In 1922 he delivered the Tarner Lectures, subsequently published as *The Mind and Its Place in Nature*, and was invited to succeed McTaggart as lecturer at Cambridge. After McTaggart's death in 1925, he oversaw the publication of the second volume of McTaggart's *The Nature of Existence*, which served as the stimulus for writing his own monumental *Examination of McTaggart's Philosophy*. (This is the book of choice for any metaphysician who is sentenced to exile on a desert island.) From 1933 until his retirement in 1953 he was Knightbridge Professor of Moral Philosophy at Cambridge. His other books include *Five Types of Ethical Theory* (1930) and two collections of papers, *Ethics and the History of Philosophy* (1952) and *Religion, Philosophy, and Psychical Research* (1953). After his death his student Casimir Lewy published his courses of lectures on Leibniz and Kant.

The scope of Broad's interests was vast. Selected for attention in this article are four main topics: his conception of "critical philosophy," his writings on sensa and perception, his philosophy of time, and his views on the relation of mind to matter.

Not covered here are Broad's important contributions to the following areas: inductive logic (he sought to identify and justify some principle about the world that, if true, would make induction legitimate); determinism and freedom (he argued that the notion of "obligability" or moral responsibility is incompatible both with determinism and with indeterminism, making it a problematic concept); the relevance of psychic research to philosophy (he assessed the evidence for paranormal phenomena and identified metaphysical principles that would have to be given up if the reality of such phenomena

became established); and ethics. His *Five Types of Ethical Theory*, a study of the ethical systems of Spinoza, Butler, Hume, Kant, and Sidgwick, was a widely used text; he also wrote influential papers in metaethics, clarifying the status of non-naturalist intuitionism of the sort espoused by Moore, naturalist theories of the "moral sense" variety, and non-cognitivist or "interjectional" theories.

Critical versus speculative philosophy

In a discussion of the nature and value of philosophy in the introduction to *Scientific Thought*, Broad distinguished two branches of his subject, critical philosophy and speculative philosophy. The first task of critical philosophy is "to take the concepts that we daily use in common life and science, to analyse them, and thus to determine their precise meanings and their mutual relations." Concepts ripe for such analysis include the concepts of substance, cause, place, date, duty, and self. The second task of critical philosophy is to test the beliefs that we constantly assume in everyday life and science, "resolutely and honestly exposing them to every objection that one can think of oneself or find in the writings of others." Beliefs subject to such critical scrutiny include the beliefs that we live in a world of objects that are independent of our knowledge of them and that every event has a cause. We may emerge from critical philosophy with verbally the same beliefs we started with, but the process will have "enabled us to replace a vague belief by a clear and analysed one, and a merely instinctive belief by one that has passed through the fire of criticism." He then went on to characterize speculative philosophy as follows:

> Its object is to take over the results of the various sciences, to add to them the results of the religious and ethical experiences of mankind, and then to reflect upon the whole. The hope is that, by this means, we may be able to reach some general conclusions as to the nature of the Universe, and as to our position and prospects in it.

Broad noted that speculative philosophy is less certain in its results than critical philosophy, and that it must be augmented by critical philosophy if it is to be of any value. He engaged in both varieties of philosophy himself, but his strong suit was critical philosophy. I think it is fair to say that many analytic philosophers would cite Broad's definition of critical philosophy as an excellent description of what they do and Broad himself as an outstanding practitioner of it.

Sense-data and perception

Broad was one of the leading exponents of a sense-datum theory of perception. The term "sense-datum" was introduced by Russell and Moore; Broad himself almost invariably preferred the term "sensum." Though sometimes used broadly to cover the sensuous aspect of experience in general (however it may be analyzed), the terms "sense datum" and "sensum" have for Broad and other philosophers of his era a narrower and more precise meaning. The notion of a sensum has application only if one adopts an act–object analysis of sensory experience. To see what this means, consider the various types of sensory experience arranged in an order, starting with those

of sight, passing through those of hearing, taste, and smell, and ending with bodily sensations like headache. At the beginning of the series, Broad claimed, it seems plausible to analyze a sensation of red into two components, an act of sensing and a red object. At the other end of the series, it does not seem plausible to analyze a sensation of headache into an act of sensing and a "headachy" object. Having a headache is not sensing some*thing* – it is sensing some*how*. In the middle of the series (with taste and smell), it may not be obvious whether one can distinguish act and object. Some philosophers have assimilated the entire series to one or the other end of it, advocating either an act–object analysis across the board (H. H. Price) or an objectless "way of sensing" analysis across the board (Thomas Reid). Broad saw no reason to treat the entire series uniformly; he took bodily sensations to be objectless, but espoused an act–object analysis at least for sight, hearing, and touch. "It seems to me much more certain that, in a sensation of red, I *can* distinguish the red patch and the act of sensing it, than that, in a sensation of headache, I *cannot* distinguish a headachy object, and an act of sensing it" (1923: 256). The red patch that figures as the object-component in the sensation of red is the sort of thing Broad meant by a sensum.

That is not to say that when I am seeing a ripe tomato, the tomato is a sensum. Even if my experience of a tomato were a total hallucination and there were no red physical objects in my environment, there would still be something red that I am sensing, and that something is a sensum. Thus sensa are not automatically to be identified with physical objects or even parts of them; their relation to physical things is a more complex affair to be discussed further below.

The theory of sensa may be expounded further by noting some of the familiar facts it is meant to explain. When viewing a penny or a coffee cup tilted away from my line of sight, I may be certain that I am having the experience expressed by "This looks elliptical to me," even though I know that in fact the penny or the cup is not elliptical but round. This much is supposed to be a fact on which all parties agree. The sensum theory analyzes the situation as involving "the actual existence of an elliptical object, which stands in a certain cognitive relation to me on the one hand, and in another relation, yet to be determined, to the round penny" (1923: 237–8). This elliptical object is a sensum. Broad pithily conveyed the guiding motivation for positing it as follows: "If, in fact, nothing elliptical is before my mind, it is very hard to understand why the penny should seem *elliptical* rather than of any other shape" (p. 240).

Generalizing from what Broad says about the penny, we may put the essential core of the sensum theory as follows: whenever any object x appears to a subject S to have a property F, it does so because S is directly aware of an item y (a sensum) that really does have the property F. The item y is the sensum, and its relation to x cannot in general be identity (since if x appears F without being F, y is F and x is not). (Certain restrictions are to be understood as attaching to this formulation; "appears F" is used phenomenally, not comparatively, and the variable F ranges over color, shape, and distance.) It is generally held that sensa themselves, unlike physical objects, never appear to have any property F without really having it. This is implicit in the reason for positing sensa in the first place: if sensa could appear to have properties they do not really have, we would have to posit a second tier of sensa to be the bearers of the properties apparently possessed by sensa in the first tier.

What about the converse assumption, that sensa have *only* the properties they appear to have? Broad denied this, holding that sensa may be more variegated or determinate than they appear to be. This enabled him to avoid the objection that sensa would be indeterminate in their properties. If a sensum appears to be many-speckled without appearing to be exactly *n*-speckled for some *n*, Broad does not have to say that the sensum has speckles without having any definite number of them.

Many writers assume that visual sensa must have only two dimensions – that they are extended in length and breadth, but altogether lacking in depth. Broad argued to the contrary that visual sense data are as fully voluminous or three-dimensional as any objects in physical space. Many writers assume that sensa are mental entities. This, too, Broad denied. According to him, they are neither mental nor physical, but have a leg in each realm (1925: 184). For example, they are like physical objects in having spatial qualities like extension and shape, but like mental things in being private to observers and sense modalities. Their privacy, however, does not mean that sensa are existentially mind-dependent; like Russell, Broad accepted it as a real possibility that there can be unsensed sensibilia.

Sense-data have been out of vogue for nearly fifty years. Opposition to them has stemmed from two main motives. First, they are hard to accommodate within a purely physicalist view of the universe: if the experience of red, whether veridical or not, involves a literally red object, it is hard to see with what brain entities or processes this red object could be identified. Second, sensa make difficulties for direct realist accounts of perception: they are often thought to constitute a "veil" between perceivers and the physical world, cutting us off not only from direct perception of physical things but knowledge of them as well.

What, then, are the alternatives to admitting sensa? A radical alternative is to deny (with Daniel Dennett and others) that there is a sensuous element in experience at all, in which case there would be nothing for the sensum theory to analyze. Broad would have dismissed this suggestion as a flagrant denial of the facts. He did, however, recognize two alternative analyses of the facts in addition to the sensum theory: the multiple relation theory and the multiple inherence theory. The first of these alternatives is mentioned, though not discussed, in *Scientific Thought*; both are discussed in *The Mind and Its Place in Nature*.

One way to understand the differences among the three theories is to see what each would say about the phenomenon of perceptual relativity: the fact that the same object can appear to have different properties to different observers or from different viewpoints, as when water feels hot to one hand and cold to another, or a mountain looks blue from a distance and green close up. It would be contradictory, of course, to say that the same mountain is both green and blue, period. But there are three ways to state the facts of perceptual relativity without contradiction. First, we can say that the incompatible colors inhere in different subjects. This is what the sensum theory says: I sense one batch of sensa (blue ones) when I am viewing the mountain from afar and another batch of sensa (green ones) when I am standing on the summit. Second, we can say that the mountain looks blue as I approach it on the highway, that it looks green when I get there, and that on at least one of these occasions it looks to have a color that nothing in the situation actually has. This is what Broad called the "multiple relation theory of appearing"; it holds that appearing *F* is an unanalyzable relation between an

object, a property, and a mind, not involving the existence of any entity that really is *F*. Third, we can say that the mountain *is* blue *from here* and that it *is* green *from there*, avoiding contradiction by expanding the number of places in the relation of inherence. This is the multiple inherence theory, in which we give up the ordinary view that the inherence of a color in a thing is a two-term relation between the color and the thing; rather, it is a three-term relation between a color, a thing, and a place or a viewpoint. Colors do not inhere in objects simply, but only in objects (or "regions of pervasion") from places (or "regions of projection").

The multiple relation theory has the counterintuitive consequence that objects "can have qualities which are different from and inconsistent with those which they seem on careful inspection to have" (1925: 160). The multiple inherence theory involves a puzzling new form of inherence; in addition, it has the puzzling consequence that the colors of objects are "causally adventitious" to them, in the sense that the immediate causal determinants of the color pervading a region lie not in that region but in some other region, a "region of projection" containing a suitably functioning brain. Broad did not think either of these theories was decisively refutable, but he found the sensum theory preferable on the whole.

A further possibility is worth mentioning. One may accept the verbal formula Broad uses in characterizing the multiple relation theory – "an object can appear *F* without anything's being *F*" – without taking the relation of appearing *F* as unanalyzable. That, in effect, is what Roderick Chisholm does, analyzing "*x* appears *F* to *S*" as "*x* causes *S* to sense *F*-ly." He abandons the act–object analysis of sensing in favor of an adverbial approach, according to which to have a sensation of red is simply to sense in a certain way. He then analyzes the relation of an object's looking red to *S* as a matter of the object's causing *S* to sense redly.

It remains to say something about Broad's views on the relation of sensing to perceiving and of sensa to physical objects. When I perceive something, I do not merely sense a sensum; I also believe in an object (e.g. a bell or a candle) to which the sensum is related. Broad devoted considerable attention to analyzing this belief and its object. He worked out elaborate answers to questions like these: how do physical objects cause sensa, and how are the places, dates, durations, shapes, and sizes of physical objects to be defined or known in terms of the corresponding features of sensa? The corresponding features of sensa that go by the same name are sometimes literally the same and sometimes not. Sensa and physical objects both have shape in the same sense, but they do not have location in the same sense. Sensa are literally located only in their own spaces (e.g. a sensum of color may be in the center of one's visual field). They may also be assigned locations in physical space, but only in a "Pickwickian" sense. Roughly, to say that a visual sensum *s* is "in" physical place *p* means this: if I turn my head to bring *s* into the center of my visual field and then follow my nose, I will bring myself closer and closer to *p*, obtaining along the way a series of sensa like *s* but becoming larger and brighter until I eventually advance beyond *p* and the *s*-like sensa disappear.

Broad was never a phenomenalist, one who believes that physical objects are composed (or logically constructed) entirely of sensa. He believed that physical objects are heterogeneous composites, containing as literal parts atoms or whatever tinier particles are recognized by the best science of the day and containing as Pickwickian parts sensa belonging to the various sense realms. He also espoused something like the

traditional distinction between primary and secondary qualities, maintaining that shapes inhere literally both in the scientific constituents of physical objects and in sensa, while colors inhere in sensa alone. His main reason for denying that colors inhere in physical objects was that we need to refer to the shapes of physical objects to explain why we sense sensa of various shapes, but do not need to assign colors to physical objects in order to explain why we sense colored sensa.

Philosophy of time

Broad had a good deal to say about the nature of space and time, including interpretations in *Scientific Thought* of Einstein's Special and General Theories of Relativity, which were then fairly new on the scene. I focus here on his more purely metaphysical views about time, as presented both in *Scientific Thought* and *Examination of McTaggart's Philosophy.*

Some philosophers hold that only the present is real; others hold that past, present, and future are all equally real. In *ST*, Broad advanced a theory intermediate between these two, accepting the reality of the present and the past, but holding that "the future is simply nothing at all" (1923: 66). The time series is like a growing line, and it possesses a direction because "fresh slices of existence" are always being added to the forward end of it. He drew from this the conclusion that judgments ostensibly about future events are neither true nor false at the times when they are made, since there is nothing then in existence to make them true or false (p. 73).

Broad distinguished two aspects of time or of temporal facts, which he called the "extensive" (or static) and the "transitory" (or dynamic) aspects. The distinction is closely related to McTaggart's distinction between the A series and the B series. Call the relations of being earlier than, later than, and simultaneous with "B relations"; call the characteristics of pastness, presentness, and futurity "A-characteristics." An A series is then any series of events or moments whose members have A-characteristics, and a B series any series whose members are related by B-relations. McTaggart noted that truths involving the B-relations are permanent, while truths involving the A-characteristics are transitory. An event that is earlier than another event is always earlier than it, but an event that is future will not always be future: it will become less and less remotely future, then it will become present, and finally it will become more and more past.

A great divide in philosophies of time separates those who acknowledge the transitory aspect of time and those who reject it. Russell and many others deny it, affirming that temporal facts are exhausted by those involving the B-relations. Broad upheld it, agreeing with McTaggart that the transitory aspect of time is essential to it. He did not, however, believe that events become present in the way that may be suggested by McTaggart's language, that is, the events are already strung out and become present as the palings of a fence become illuminated by the passage of a spotlight. Becoming is not analogous to qualitative change, in which a subject that already exists acquires a new property; rather, to become present is just to "become," in an absolute sense. Broad's adherence to the transitory aspect of time is reflected instead in his insistence on the indispensability of tense, for tensed statements are precisely those that may change truth value with the passage of time.

The indispensability of tense – the thesis that tensed verbs cannot be done away with in the analysis of temporal discourse – is perhaps Broad's most important thesis in his later philosophy of time. He opposed both Russell's analysis of tense in terms of tenseless copulas and B-relations and McTaggart's analysis of tense in terms of tenseless copulas and A-characteristics. According to Russell, an utterance of the sentence-type "it is now raining" means that an occurrence of rain is simultaneous with that very utterance; in analogous fashion, an utterance of "it has rained" or "it will rain" would mean that an occurrence of rain is earlier or later than that very utterance. Broad expressed doubt about whether the kind of self-reference involved here is really possible and about whether tenseless verbs are anything but a philosopher's fiction. His main objection, however, was simply that Russell's analysis leaves out the transitory aspect of temporal facts. If an occurrence of rain is (tenselessly) simultaneous with a certain utterance, it is always simultaneous with that utterance, making any utterance of "it is now raining" true eternally if it is true at all.

McTaggart's presupposition that tense is eliminable is an essential part of his notorious argument for the the unreality of time, an argument that Broad subjected to penetrating analysis. McTaggart, unlike Russell, believed that there could not be time without an A-series: a series of events or moments exemplifying the characteristics of past, present, and future. His case against the reality of time was that the A-series involves a contradiction: the A-characteristics are mutually incompatible, yet each item in any A-series must have them all. To this the obvious objection is that each event has all of the A-characteristics only successively, and there is no contradiction in that. An event that is now present is not *now* past and future; rather, it *has been* future and *will be* past. But McTaggart anticipated this objection, and replied that our attempt to remove the contradiction only raises it anew. When we say that *S* has been (will be, is now) *P*, we are saying that *S* is *P* at a moment of past (future, present) time. Thus to say that an event has been future, is now present, and will be past implies that there is an A-series of moments. And this, McTaggart alleged, brings back a contradiction just like the original one: every moment, like every event, is past, present, and future.

But why did McTaggart think there was a contradiction to begin with in saying that an event is future, present, and past, a contradiction that remains even if we add the qualification "successively"? To say that an event is successively future, present, and past is to say (if it is now present) that it was future and will be past. According to Broad, it is at this point in the argument that McTaggart's assumption that tense is eliminable plays a crucial role. Broad articulated the assumption as follows: what is meant by a sentence with a tensed verb or copula must be completely and more accurately expressible by a sentence in which there is no tensed verb or copula, but only temporal predicates and tenseless verbs or copulas. To highlight the fact that the more accurate expression must be free of tense, let us use "be" as a tenseless copula. Then McTaggart's claim is (e.g.) that "*e* was future" means "for some moment *m*, *e* be future at *m* & *m* be past." Well, if m *be* past, it is timelessly or sempiternally past. And that contradicts the assumption, inherent in belief in the A-series, that every moment is sometimes future and sometimes present. Thus Broad concludes:

> [T]he source of McTaggart's regress is that, if you take the "is" in "t is present" to be timeless, you will have to admit that t is also past and future in the same timeless sense of "is".

> Now this is impossible, for it is obvious that t can have these predicates only in succession. If, to avoid, this, you say that the "is" in "t is present" means "is now", you have not got rid of temporal copulas. Therefore, if you are committed at all costs to getting rid of them, you will not be able to rest at this stage. (1933: 314–15)

So Broad insisted on the ineliminability of tense. Russell's attempt to eliminate tense in favor of the B-relations ignores the transitory aspect of time, and McTaggart's way of getting rid of it makes the transitory aspect contradictory. The moral Broad drew is that if we wish to do justice to the transitory aspect of time, we must take tense seriously.

Broad also discussed at length the ontological categories of thing and process. One of the differences is that things endure literally through time, whereas a process persists only in virtue of having distinct parts or phases that exist at various moments within the interval. If I say "This is the same chair I sat in yesterday," I mean that literally the same object I sat in yesterday is here now, but if I say "I am still hearing the same buzzing noise," I mean only that I am hearing later phases of a process whose earlier phases I heard before. Those who believe in the dynamic aspect of time commonly hold that identity through time is a matter of thing-like endurance, while those who embrace a static concept of time typically hold that identity through time is really a matter of process-like persistence. Confounding expectations on this score, Broad combined his belief in the transitory aspect of time with the view that things are dispensable in ontology in favor of logical constructions out of processes. As he sometimes bluntly put it, a thing is just a long and boring event.

Mind and matter

In the concluding chapter of *Mind and Its Place in Nature*, Broad undertook to classify the various possible metaphysical theories on the relation of mind to matter, to sum up their strong and weak points, and to decide between them. His scheme of classification yielded seventeen types of possible theory, which he thought could be narrowed down to three or four best options and one that was most reasonable overall. I now give a somewhat simplified description of Broad's scheme and of his own favored alternative, which he called "emergent materialism."

Suppose the X-properties of anything follow with logical or metaphysical necessity from some selection of its Y-properties (or the Y-properties of its parts). This could happen because the X-properties are identical with the Y-properties or are analyzable in terms of them. In this case, X-properties are *reducible* to Y-properties.

Suppose the X-properties of a thing are not reducible to the Y-properties of its parts or the relations among them, but do follow with nomological necessity from these Y-properties and relations. In this case, X-properties are *emergent* from the Y-properties.

With these preliminary notions granted, we can define Broad's notion of a "differentiating attribute" (or for short, simply an attribute): an attribute is a highly general property that is instantiated in the universe without being either reducible to or emergent from properties of any other type.

Broad analyzed materiality as the conjunction of extension, publicity, persistence, and existential independence from observing minds. He analyzed mentality as a

hierarchy of properties ranging from bare sentience up through the higher cognitive (both intuitive and discursive) and affective capacities.

We can now arrive at most of the positions in Broad's scheme by asking the following questions about each of materiality and mentality: Is it instantiated in the universe or not? If so, is it reducible, emergent, or neither? And if it is reducible or emergent, with respect to what other properties is it emergent or reducible?

If materiality and mentality are both instantiated in the universe, but neither is reducible to or emergent from anything else, that makes both of them attributes in Broad's sense, giving us the position he called "dualism." He subdivided this according to whether materiality and mentality can or cannot inhere in the same substance. If the answer is yes, we have dualism of compatibles, the position of Spinoza; if it is no, we have dualism of incompatibles, the position of Descartes.

If materiality is an attribute but mentality is not, we have the family of theories Broad called "materialist." This subdivides according to the three ways in which mentality might fail to be an attribute. If mentality is not instantiated in the universe at all, we have pure materialism (or what would nowadays be called eliminative materialism). If mentality is reducible to something else (determinates of materiality, presumably), we have reductive materialism. Broad discussed two chief varieties of this, "molar behaviorism," according to which having a mental state just means behaving in certain ways, and "molecular behaviorism," according to which mental processes are to be identified with processes in the brain and nervous system. Finally, if mentality is emergent, we have emergent materialism, according to which mental properties emerge as novel properties of material systems that achieve a certain degree of complexity.

If mentality is an attribute but materiality is not, we get the family of theories Broad called mentalist. As with materialism, there are three possible varieties: pure mentalism, reductive mentalism, and emergent mentalism. The actual mentalists Broad mentions – for example Berkeley, Leibniz, and McTaggart – are all of the pure variety. It might be thought that phenomenalism affords an example of reductive mentalism, but most phenomenalists turn out to be either pure mentalists (because they hold that nothing in the universe exemplifies *all* the traits requisite for materiality) or neutralists (because they reduce matter to properties of "neutral" sense data in the manner to be described next).

Finally, if neither materiality nor mentality qualifies as an attribute, we get the family of theories Broad classified as neutralist. Somewhat extravagantly, neutralism admits of nine subdivisions. Broad singled out two forms of neutralism as especially worthy of attention. First, there is the view of Samuel Alexander in *Space, Time, and Deity* that mind and matter both emerge from purely spatiotemporal attributes. Second, there is the view of Russell in *The Analysis of Mind* that materiality is not strictly instantiated at all (even though its various requisites are separately instantiated) and that mentality is either reducible to or emergent from properties of sense data that are themselves neither mental nor physical.

Broad went on to argue that many of the seventeen types of theory can be quite definitely ruled out. Pure materialism and the three varieties of neutralism that say mentality is not instantiated can be eliminated immediately, he claimed, for mentality at least *seems* to be instantiated, and if there are seemings, there are events that instantiate mentality. He believed that both varieties of reductive materialism could also be

ruled out. Against molar behaviorism, he pointed out that many of the mental states we observe within ourselves are not identical with their associated patterns of behavior, and he also raised the doubt whether every mental state even has a pattern of behavior coextensive with it. Against molecular behaviorism, he raised the objection that neural processes have properties (e.g. taking place swiftly or slowly) that do not apply to having a sensation of red.

It remains to say something about the position Broad judged most reasonable on the whole, namely, emergent materialism. Because it implies that mental properties are not reducible to physical properties, this is a form of what is sometimes called property dualism, even though not a form of dualism in Broad's own sense (which requires that mentality be an attribute). The idea is that mental properties begin to be displayed when matter reaches a certain level of complexity. They are dependent on matter for their instantiation and are wholly determined, causally or nomically speaking, by material configurations. However, the mental properties of an organism "*could* not, even in theory, be deduced from the most complete knowledge of the behavior of its components, taken separately or in other combinations, and of their proportions and arrangements in this whole" (1925: 59). In this respect, Broad believed mental properties to be like the properties of chemical compounds and unlike the properties of clocks. Someone who had never seen a clock before could predict its behavior from the laws of physics together with knowledge of the clock's parts and how they are put together. By contrast, someone who had never examined common salt before could not predict its properties from the laws of physics together with complete knowledge of the properties of sodium and chlorine (taken separately and in other combinations) and how they are put together in the new compound.

There are, of course, psychophysical laws relating mental properties to the physical properties of their bearers. But Broad believed these laws to be ultimate "trans-ordinal" laws: laws not deducible from laws already known to hold at the lower level, but discoverable instead only after we have become acquainted with objects and properties at the higher level. He conceded that the properties of chemical compounds, which he used as examples of emergent properties, might turn out with the growth of our physical knowledge or mathematical competence not to be emergent after all. But he thought that the traditional secondary qualities (whether conceived of naively as intrinsic properties of external things or in more sophisticated fashion as appearances to perceivers) were inherently emergent and that the laws connecting their instantiation with properties of microphysics would necessarily be of the trans-ordinal type. Not even a mathematical archangel with microscopical powers of perception, Broad ventured to assert, would be able to predict that ammonia smells acrid or that the sky looks blue.

Bibliography

Works by Broad

1914: *Perception, Physics, and Reality*, Cambridge: Cambridge University Press.
1923: *Scientific Thought*, London: Kegan Paul, Trench, Trubner & Co.
1925: *The Mind and Its Place in Nature*, London: Kegan Paul, Trench, Trubner & Co.
1930: *Five Types of Ethical Theory*, London: Kegan Paul, Trench, Trubner & Co.

1933: *Examination of McTaggart's Philosophy*, vol. I, Cambridge: Cambridge University Press.
1938: *Examination of McTaggart's Philosophy*, vol. II, Cambridge: Cambridge University Press.
1952a: *Ethics and the History of Philosophy*, London: Routledge and Kegan Paul.
1952b: *Religion, Philosophy, and Psychical Research*, London: Routledge and Kegan Paul.
1975: *Leibniz: An Introduction*, ed. C. Lewy, Cambridge: Cambridge University Press.
1978: *Kant: An Introduction*, ed. C. Lewy, Cambridge: Cambridge University Press.

Works by other authors

McLaughlin, Brian (1992) "The Rise and Fall of British Emergentism," in *Emergence or Reduction?*, ed. A. Beckerman, H. Flohr, and J. Kim, Berlin: De Gruyter.

Schilpp, P. A. (ed.) (1959) "The Philosophy of C. D. Broad," vol. X in *The Library of Living Philosophers*, New York: Tudor. (This volume contains Broad's autobiography, critical essays on Broad's philosophy by twenty-one authors, Broad's responses to them, and a complete bibliography of his writings.)

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

5

Ludwig Wittgenstein (1889–1951)

P. M. S. HACKER

Background

Ludwig Josef Johann Wittgenstein dominates the history of twentieth-century analytic philosophy somewhat as Picasso dominates the history of twentieth-century art. He did not so much create a "school," but rather changed the philosophical landscape – not once, but twice. And his successors, within the broad stream of analytic philosophy, whether they followed the paths he pioneered or not, had to reorient themselves by reference to new landmarks consequent upon his work. He completed two diametrically opposed philosophical masterpieces, the *Tractatus Logico-Philosophicus* (1921) and the *Philosophical Investigations* (1953). Each gave rise to distinct phases in the history of the analytic movement. The *Tractatus* was a source of Cambridge analysis of the interwar years, and the main source of the logical positivism of the Vienna Circle. The *Investigations* was a primary inspiration for the form of analytic philosophy that flourished in the quarter of a century after the end of the Second World War, with its center at Oxford and its circumference everywhere in the English-speaking world and beyond. He taught at Cambridge from 1930 until his premature retirement in 1947. Many of his pupils became leading figures in the next generation of philosophers, transmitting his ideas to their students.[1]

Wittgenstein's central preoccupations at the beginning of his philosophical career were with the nature of thought and linguistic representation, of logic and necessity, and of philosophy itself. These themes continue in his later philosophy, from 1929 onwards, although philosophy of mathematics occupied him intensively until 1944 and philosophy of psychology increasingly dominated his thought from the late 1930s until his death. Having been trained as an engineer, he came to Cambridge in 1911, without any formal education in philosophy, to work with Russell. He was poorly read in the history of the subject, and intentionally remained so in later years, preferring not to be influenced by others. He had read Schopenhauer in his youth, and traces of *The World as Will and Representation* are detectable in the *Tractatus* discussion of the self and the will. He acknowledged the early influence upon him of the philosopher-scientists Boltzmann (in particular, apparently, of his *Populäre Schriften*) and Hertz (especially his introduction to *The Principles of Mechanics*). Apart from these figures, the main stimuli to his thoughts were the writings of Frege and Russell on logic and the

foundations of mathematics. In later years, as he put it, he "manufactured his own oxygen." He certainly read some Kant when he was prisoner of war in Cassino, some of the works of Augustine, Nietzsche, Kierkegaard, and Plato, but did not cite these as influences upon him.[2] The only later influences he acknowledged were Oswald Spengler, and discussions with his friends Frank Ramsey and Piero Sraffa.

His style of thought and writing were idiosyncratic. He was able to dig down to the most fundamental, and typically unnoticed, presuppositions of thought in a given domain. Where philosophers had presented opposing views of a topic, and debate had long continued polarized between alternatives, for example between idealism and realism in epistemology, or dualism and behaviorism in philosophy of mind, or Platonism and intuitionism in philosophy of mathematics, Wittgenstein did not side with one or another of the received options, but strove to find the agreed presuppositions common to both sides of the venerable dispute, and then challenged these. His insights were typically written down in highly condensed form: often a single sentence, a brief paragraph, or a fragment of an imaginary dialogue. Writing standard consecutive prose distorted his thoughts, and, for the whole of his life, his writings were sequences of remarks, entered into notebooks, from which he later extracted and ordered the best. This, together with his great gift of simplicity of style, rich in metaphor, simile, and illuminating example, gives his philosophical writing power and fascination, as well as formidable interpretative difficulty. In one sense, he had the mind of an aphorist, for what is visible on the page is often no more than the trajectory of a thought, which the reader is required to follow through. No other philosopher in the history of the subject shared his cast of mind or style of thinking. The closest in spirit are the philosophically-minded aphorists Georg Christoph Lichtenberg (whom he much admired) and Joseph Joubert (with whose writings, it seems, he was not acquainted). During his lifetime he published only one book, the *Tractatus*, and one article "Some Remarks on Logical Form," written for the Mind and Aristotelian Society meeting in 1929. By the time of his death, he had more or less completed the *Investigations* (Part 1), and wished it to be published. As for the rest, he left to his literary executors the decision on what parts of his literary remains of more than twenty thousand pages of notes and typescripts should be published.[3]

After the posthumous publication of the *Philosophical Investigations* in 1953, his literary executors edited numerous volumes of his unfinished typescripts and notes from all phases of his philosophical career. *Notebooks 1914–1916* consists of preparatory notes for the *Tractatus*. *Philosophical Remarks* was written in 1929, and represents the stage at which the philosophy of the *Tractatus* was starting to crumble. *Philosophical Grammar* is an editorial compilation from typescripts written in the years 1931–4, and signals the transformation of Wittgenstein's thought, abandoning the philosophy of the *Tractatus* and articulating his new methods and ideas. Half of it concerns problems in the philosophy of mathematics, a subject which was at the center of his interests from 1929 until 1944. *The Blue and Brown Books* consists of dictations to his pupils, given in 1933–5. It elaborates his new philosophical methods and his transformed conception of philosophy, and examines problems in the philosophy of language, epistemology, metaphysics, and philosophy of psychology. The *Remarks on the Foundations of Mathematics* is a selection from typescripts and manuscripts written between 1937 and 1944. *Zettel* is a collection of cuttings Wittgenstein himself made

from typescripts written between 1929 and 1947, although most of the remarks date from the period 1944–7. The themes are mainly topics in the philosophy of language and philosophy of mind. The four volumes of *Remarks on the Philosophy of Psychology* and *Last Writings on the Philosophy of Psychology* are notes written between 1947 and 1951. *On Certainty* and *Remarks on Colour* were written at the very end of his life, the former being unique among his works in its exclusive focus on epistemological themes. Apart from other minor writings, for example on Frazer's *Golden Bough* or aphorisms and general cultural observations jotted down amidst his philosophical reflections and gathered together in *Culture and Value*, five volumes of lecture notes taken by his students have been published. The complete *Nachlass* is currently being published in electronic form.

Wittgenstein is unique in the history of philosophy as the progenitor of two profoundly opposed comprehensive philosophies. To be sure, there are continuities of theme between the two: the nature of linguistic representation, of logic and laws of thought, of the relation between thought and its linguistic expression, of the intentionality of thought and language, of metaphysics and of philosophy itself are topics examined in detail in the *Tractatus* and then re-examined in the later philosophy. There are also continuities of philosophical judgment. Many of the negative claims in the *Tractatus* are reaffirmed in the later works, in particular his criticisms of Frege and Russell, his denial that philosophy can be a cognitive discipline, his rejection of psychologism in logic and of logicism in the philosophy of mathematics. And many of the fundamental insights that informed the *Tractatus*, for example that there is an internal relation between a proposition and the fact that makes it true, that the propositions of logic are senseless but internally related to inference rules, that the logical connectives and quantifiers are not function names, that ordinary language is in good logical order, are retained in the later philosophy. Nevertheless, the insights that are thus retained undergo transformation, are relocated in the web of our conceptual scheme, are differently elucidated, and quite different consequences are derived from them. In general, the two philosophies represent fundamentally different philosophical methods and ways of viewing things. The *Tractatus* is inspired and driven by a single unifying vision. It was intended to be the culmination and closure of the great essentialist metaphysical tradition of western philosophy. An insight into the essential nature of the elementary proposition was held to yield a comprehensive account of the nature of logic and of the metaphysical form of the world, the nature and limits of thought and language. An ineffable metaphysics of symbolism was wedded to an equally ineffable solipsistic metaphysics of experience and to an atomist, realist, ontology.

The *Tractatus*

The two major thinkers whose work both inspired Wittgenstein and constituted the main target of his criticisms were Frege and Russell. They had revolutionized logic, displacing the subject/predicate logic of traditional syllogistic by the function theoretic logic based on the generalization of the mathematical theory of functions. Frege had invented the logic of generality, the predicate calculus (see FREGE). Both philosophers repudiated psychologism in logic and idealism in metaphysics and epistemology, propounding instead forms of realism. Both had tried to demonstrate the reducibility

of arithmetic to pure logic, Frege in *The Basic Laws of Arithmetic* (1893, 1903) and Russell, together with Whitehead, in *Principia Mathematica* (1910). It was, above all, their conception of logic that set the agenda for the young Wittgenstein.

Frege and Russell thought that logic was a science with a subject matter. The propositions of logic, they held, are characterized by their absolute generality. On Frege's view they are perfectly general propositions concerning sempiternal relations between thoughts (propositions), articulating laws of truth valid for all thinking. According to Russell, logic is the science of the perfectly general. Its propositions are descriptions of the most general facts in the universe. Hence neither would have considered a simple tautology such as "Either it is raining or it is not raining" as a proposition of logic, but would have conceived of it as an instantiation of the logical proposition "$(p)\ (p \vee \sim p)$." Both tended to view rules of inference ("laws of thinking") as related to the propositions of logic ("laws of truth") somewhat as technical norms specifying a means to an end are related to laws or regularities of nature. The laws of truth according to Frege describe the immutable relations between thoughts (propositions) irrespective of their subject matter; according to Russell, they are the most general laws governing the facts of which the universe consists. Accordingly, rules of inference are technical norms, dependent on such general laws, ensuring that if one wishes to think correctly, i.e. infer only truths from truths, one will do so. The logical systems the two philosophers had invented were axiomatized, and they viewed the axioms as self-evident truths. Frege conceived of thoughts and of the two truth-values as logical objects, and of the notions of object, concept, first- and second-level function as ultimate *summa genera*, drawing ontological distinctions "founded deep in the nature of things." The logical connectives he thought to be names of logical entities, unary or binary first-level functions mapping truth-values on to truth-values, and the quantifiers to be names of second-level functions. Russell held that terms such as "particular," "universal," "relation," "dual complex," are names of logical objects or "logical constants" signifying the pure forms which are the *summa genera* of logic, the residue from a process of generalization which has been carried out to its utmost limits. We understand such expressions, he thought, on the basis of "logical experience" or intuition. Both philosophers held natural language to be logically imperfect, containing vague and ambiguous expressions or names without reference, and hence, Frege thought, allowing the formation of sentences without a truth-value. They viewed their own notations as logically perfect languages. From the post-Wittgensteinian perspective, Frege and Russell were radically mistaken about the nature of logical truths (conceiving of them as essentially general), about the nature of logical necessity, about the content of logical truths, about the status of the axioms of logic, about the character of the logical connectives and quantifiers, and about the relation between the truths of logic and rules of inference. If we are any clearer on these matters than they, it is largely due to Wittgenstein.

In the *Tractatus*, Wittgenstein accepted some of the salient doctrines of Russell and Frege. Like them, he adopted a (different) variant of metaphysical realism in the *Tractatus* ontology of simple sempiternal objects, of complexes, and of facts. He accepted unreflectively the assumption that the fundamental role of words is to name entities (although this role was denied to logical operators and to categorial expressions) and of sentences to describe how things are in reality. He thought that there must be a connection of meaning between words and the entities they name, that language

acquires content by means of such a connection with reality. He agreed with their antipsychologism in logic. He accepted Frege's demand of determinacy of sense, although unlike Frege, he thought that the vagueness of natural language was merely superficial and analyzable into disjunctions of determinate possibilities. And, like Frege, Russell and many others, he assumed that the logical connectives and quantifiers are topic-neutral. Some of these commitments he was later to abandon, others he reinterpreted.

Unlike Frege and Russell, Wittgenstein held that ordinary language is in good logical order. For logic is a condition of sense, and insofar as sentences of ordinary language express a sense, convey thoughts, they are in good order – any appearance to the contrary (e.g. vagueness) being a feature of the surface grammar of expressions, which will disappear on analysis. Insofar as they fail to express a sense, they are ill-formed pseudo-sentences. Hence it is not the task of philosophy to devise a logically ideal language, although devising a logically perspicuous notation will enable the philosopher to lay bare the true logical forms of thoughts, which are obscured by the surface grammar of ordinary language. According to the *Tractatus* the fundamental function of language is to communicate thoughts by giving them expression in perceptible form. The role of propositions (sentences with a sense) is to describe states of affairs, which may or may not obtain. If the state of affairs depicted by a proposition obtains, then the proposition is true, otherwise it is false. Propositions are composed of expressions. Logical expressions apart, the constituent expressions in a proposition are either analyzable, definable by analytic definition or paraphrase, or unanalyzable. Unanalyzable expressions are simple names, which are representatives of simple objects. The simple objects are the meanings of the names. Hence names link language to reality, pinning the network of language on to the world. Names have a meaning only when used as representatives, and they are so used only in the context of a proposition. The elementary (logically independent) proposition is a concatenation of names in accordance with logical syntax. It does not name anything, *pace* Frege (who thought sentences name truth-values) and Russell (who thought they name complexes), but depicts a (possible) state of affairs, which is isomorphic to it given the rules of projection, and asserts its existence. The names in an elementary proposition must possess the same combinatorial possibilities in logical syntax as the metaphysical combinatorial possibilities of the objects in reality that are the constituents of the state of affairs represented. The logical syntax that underlies any possible means of representation mirrors the logico-metaphysical forms of reality. *Pace* Frege and Russell, the assertion sign has no logical significance. Unlike Frege, who thought that there were alternative analyses of propositions, and unlike Carnap, who, in the 1930s, thought that we can choose between different logics, Wittgenstein thought that analysis is unique and that in logic there are no options.

The metaphysics of the *Tractatus* was realist (as opposed to nominalist), pluralist (as opposed to monist), and atomist. The sempiternal objects that constitute the substance of all possible worlds include properties and relations of categorially distinct types. It is far from clear what kinds of things Wittgenstein had in mind, but they are arguably such items as minimally discriminable shades of color, tones, etc. as well as spatio-temporal points in the visual field. Objects are simple (this is mirrored by the logical simplicity, i.e. unanalyzability, of their names). They have internal and external properties.

Their internal properties constitute their (essential) form: their combinatorial possibilities with other objects (this is mirrored by the logico-syntactical combinatorial possibilities of their names). Different objects belonging to the same ontological category (e.g. different shades of color) have a common form (namely, color). The external properties of objects are accidental: their contingent concatenations with other objects to form actual states of affairs. A state of affairs is a possible combination of objects (e.g. that such and such a spatiotemporal point is a certain shade of such and such a color). The obtaining or non-obtaining of a state of affairs is a fact (hence there are positive and negative facts). Elementary states of affairs are "atomic" or "independent," that is, each such state of affairs may obtain or not obtain while all other elementary states of affairs that obtain remain the same. This is reflected by the logical independence of the elementary proposition, which has no entailments. The metaphysics of experience in the *Tractatus* was apparently a form of empirical realism and transcendental solipsism (cf. Kant's empirical realism and transcendental idealism). The empirical self that is studied by psychology is not an object encountered in experience, but a (Humean) collection of experiences. The metaphysical self, which is the concern of philosophy, is a limit of experience. It is the willing self, the bearer of good and evil.

Sentences are expressions of thoughts. But thought itself is a kind of language, composed of thought-constituents. The form of a thought must mirror the form of reality no less than a proposition. Natural language is necessary for the communication of thoughts but not, it seems, for thinking – which can be effected in the "language of thought." It is mental processes of thinking and meaning that inject content into the bare logico-syntactical forms of language. What pins a name on to an object in reality that is its meaning (*Bedeutung*) is an act of meaning (*meinen*) by the name of *that* object. What renders a licit concatenation of signs a living expression of a thought is the employment of the method of projection, which is thinking the sense of the sentence, i.e. *meaning* by the sentence such and such a state of affairs. Hence the intentionality of signs is derived from the (intrinsic) intentionality of thinking and meaning (*meinen*).

The *Tractatus* account of the intentionality of thought and language is informed by the insight that thought and proposition alike are internally related to the fact that makes them true. The thought or proposition that *p* would not be the thought or proposition it is were it not made true by the fact that *p* and made false by the fact that not *p*. What one thinks when one thinks truly that *p* is precisely what is the case, and not something else (such as a Fregean *Gedanke*), which stands in some relation to what is actually the case. But what one thinks when one thinks falsely that *p* is not what is the case (since what one thinks does not obtain). Yet one does not think nothing. Indeed, what one thinks is the same, no matter whether one thinks truly or falsely. The picture theory of thought and proposition provided a logico-metaphysical explanation of how it is possible to satisfy the demands consequent upon these internal relations. It attempts to explain how it is possible for a thought to determine what state of affairs in reality will make it true, how it is possible for the content of a thought to be precisely what is the case if it is true and yet to have a content even if it is false, how it is possible that one can read off from a thought, in advance of the facts, what will make it true, and how it is possible for the "mere signs" of language to be intentional, i.e. for a name to reach up to the very object itself of which the name is the name and for the

sentence to describe the very state of affairs the existence of which will make true the proposition expressed.

Every representation is a picture of a possibility. A proposition or thought is a logical picture, whose simple constituents name sempiternal objects with determinate form. There is a metaphysical harmony between language and thought on the one hand and reality on the other; for when one thinks truly that *p*, what is the case is that *p*; and when one thinks falsely that *p*, what one thinks is precisely what is *not* the case. This "pre-established harmony" is orchestrated by a metaphysics of symbolism. Only simple names can represent simple objects. Simple names have a meaning but no sense. Relations too are objects, and only relations can represent relations; hence in the proposition "aRb," it is not "R" that represents the relation that *a* stands in to *b*, but rather it is *that* "R" stands to the right of "a" and to the left of "b" (in this notation). Only facts can represent facts, and sentences – in their symbolizing capacity – are facts, which are used to describe how things are. For it is the fact that the constituent names are arranged as they are (in accordance with logical syntax) that says that things are thus and so. Sentences have a sense but no meaning.

The possible states of affairs in reality are determined by the language-independent combinatorial possibilities of objects. Every elementary proposition depicts a possible state of affairs. It is true if the possibility depicted obtains, false if it does not. It is of the essence of the proposition with a sense to be bipolar, i.e. to be capable of being true and capable of being false.[4] This mirrors the metaphysical truth that it is of the nature of states of affairs that they either obtain or fail to obtain. The sense of a proposition is its agreement and disagreement with the existence and non-existence of states of affairs. For the proposition that *p* agrees with the fact that *p* and disagrees with the fact that not-*p*. What one thinks when one thinks that *p* is a possibility, a possibility which is actualized if one's thought is true and is not if one's thought is false. Hence one can read off a proposition or thought (which is a kind of proposition) what must be the case for it to be true, and what one thinks when one thinks that *p* is precisely what is the case if one's thought is true and what is *not* the case if one's thought is false, and is the very same thought no matter whether it is true or false.

The logical connectives are not names of functions, but rather signify truth-functional *operations* on propositions. The quantifiers are construed as operators upon a propositional function (e.g. "f*x*") which is a logical prototype collecting all propositions of a certain form (whose values are all those propositions obtained by substituting a name for the variable), hence generating logical sums or products of such sets of propositions. All possible molecular propositions can be generated by truth-functional operations upon elementary propositions. Hence all logical relations are determined by truth-functional combinations of propositions. A molecular proposition *p* entails another proposition *q* if and only if the sense of *q* is contained in the sense of *p*, i.e. if the truth-grounds of *p* contain the truth-grounds of *q*. The various operators are interdefinable, and reducible to the single operation of joint negation, namely "not . . . and not . . ." Among the truth-functional combinatorial possibilities of a given number of elementary propositions, there will always be two limiting cases (1) in which the propositions are so conjoined as to be true irrespective of the truth-values of the constituent propositions and (2) false irrespective of their truth-values. The former is a tautology and the latter a contradiction. These are the propositions of logic. Since they are, respec-

tively, true and false irrespective of how things are, they are wholly without any content, and say nothing about how things are in reality. So by contrast with other molecular propositions which are true under certain conditions (i.e. for certain assignments of truth-values to their constituents) and false under others, the propositions of logic are unconditionally true or false. Hence they are said to be *senseless*, to have, as it were, zero-sense. All tautologies say the same thing, namely nothing. But different tautologies may nevertheless differ, for every tautology is a form of a proof (since every tautology can be rewritten in the form of a *modus ponens*), and different tautologies reveal different forms of proof. It is a mark of the propositions of logic, Wittgenstein held, that in a suitable notation they can be recognized as such from the symbol alone. He invented a special notation to display this, his T/F notation. Instead of writing molecular propositions by means of symbols for logical connectives, he used truth-tables as propositional signs. Here it is immediately perspicuous from the sign alone whether a proposition is a tautology, and if so, it is visibly evident that it cannot be false. It is equally evident whether one proposition follows from another, i.e. whether the truth-grounds of one contain those of another. This showed, he thought, the nature of the propositions of logic and their categorial difference from empirical propositions.

This conception of logical truth made clear how misleading was the Frege/Russell axiomatization of logic, with its appeal to self-evidence for the axioms. Their axioms were not privileged by their self-evidence. They were tautologies no less than their theorems. They were not "essentially primitive," nor were Frege's and Russell's theorems essentially derived propositions, for "all the propositions of logic are of equal status," namely tautologies that say nothing. Hence too, contrary to Frege and Russell, the propositions of logic have no sense, and describe nothing. In an important sense, the propositions of logic have no subject matter, and logic is misconstrued as the science of the most general laws of truth or of the most general facts in the universe. Consequently, the propositions of logic do not constitute the foundations for the elaboration of technical norms of thinking on the model of the relation between laws of nature and technical norms for achieving desired ends. Rather, every tautology is internally (not instrumentally) related to a rule of inference or form of proof.

The conception of logic in the *Tractatus* was still flawed. But its flaws, which Wittgenstein was later to expose, did not significantly affect the criticisms of the Fregean and Russellian conceptions of logic. According to the *Tractatus* the only (effable) necessity is logical necessity. Every well-formed proposition with a sense must be bipolar. What philosophers had hitherto conceived of as categorial (or formal) concepts, such as *object*, *property*, *relation*, *fact*, *proposition*, *color*, *number*, etc. are, Wittgenstein argued, expressions for forms, which are represented by variables, rather than by names. Hence they cannot occur in a fully analyzed proposition with a sense. One cannot say that, for example, one is a number, that red is a color, or that A is an object, for such pseudo-propositions employ a formal concept as if it were a genuine concept, and they are not bipolar. Hence such metaphysical pronouncements (which attempt to describe non-logical necessities) are nonsense – ill-formed conjunctions of signs. But what such pseudo-propositions *try to say* is actually *shown* by genuine propositions which contain number words, color names, or other names of objects. It is shown by *features* of the expressions in such propositions, namely by

the forms of the expressions – their essential combinatorial possibilities. These are represented by the variable of which the meaningful names are substitution-instances. An immediate consequence of this is that most of the propositions of the *Tractatus* which delineate the necessary forms of language and reality are nonsense. Hence Wittgenstein's penultimate remark in the book: "My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them – as steps – to climb up beyond them."

Hence too, the conception of philosophy advocated for the future is not the practice exhibited in the book. The *Tractatus* consists largely of sentences that are neither bipolar propositions nor tautologies. They attempt to describe the essence of the world, of language, and of logic, and of the essential relations between them. But this is an attempt to say the very things that cannot be said in language, but are rather shown by language. What is thus shown is indeed ineffable. Hence metaphysics, the attempt to disclose the essential natures of things, is impossible. Once the correct logical point of view has been achieved, once the world is seen aright, the task of the *Tractatus* is completed. The task of *future* philosophy is *analysis*: clarification of philosophically problematic propositions which will elucidate their logical forms or clarify why and where (in the case of putative metaphysical propositions) they fail to accord with the rules of logical grammar. Future philosophy will not be a theory, nor will it propound doctrines or attain knowledge. It will be an activity of logical clarification. Philosophy, thus conceived, is a critique of language.

The role of the *Tractatus* in the history of analytic philosophy

In six respects the *Tractatus* introduced the "linguistic turn" in philosophy. First, it set the limits of thought by setting the limits of *language*: by elucidating the boundaries between sense and nonsense. This put language, its forms and structures, at the center of philosophical investigation. Second, the positive task for future philosophy was the logico-linguistic analysis of *sentences*. The logical clarification of thoughts is to proceed by the clarification of propositions – sentences with a sense. Third, the negative task of future philosophy was to demonstrate the illegitimacy of metaphysical assertions by clarifying the ways in which attempts to say what is shown by *language* transgresses the bounds of sense. Fourth, the *Tractatus* attempted to clarify the essential nature of the propositional *sign* by elucidating the general propositional form, that is, by giving "a description of the propositions of *any* sign-language *whatsoever* in such a way that every possible sense can be expressed by a symbol satisfying the description, and every symbol satisfying the description can express a sense, provided that the meanings of names are suitably chosen." Fifth, the logical investigation of phenomena, the unfolding of their logical forms, which was not undertaken in the book, is to be effected by logical analysis of the *linguistic descriptions* of the phenomena. (The first moves in carrying out this task were taken in the 1929 paper "Some Remarks on Logical Form," whereupon the whole project collapsed.) For the logical syntax of language is and must be isomorphic with the logico-metaphysical forms of the world. Sixth, the greatest achievement of the book, as seen by the Vienna Circle, was its elucidation of the nature

of logical necessity. This was patently made by an investigation of *symbolism*. That one can recognize the truth of a logical proposition from the symbol alone was held to contain in itself the whole philosophy of logic.

Many of these claims were later to be repudiated. But they heralded the linguistic turn, which was executed by the Vienna Circle, and, in a different way, by Wittgenstein himself in his later philosophy, and by Oxford analytic philosophy. The *Tractatus* was a paradigm of analytic philosophy in its heroic or classic phase in the interwar years. It was the major inspiration of Cambridge analysis and of logical positivism. Its program, as understood both in Cambridge and in Vienna, committed one to the method of logico-linguistic analysis of complex expressions into their simple unanalyzable constituents. It encouraged the program of reductive analysis and its mirror image, logical construction. It cleaved to the thesis of extensionality, holding all non-extensional contexts to be either eliminable, merely apparent, or illicit. It repudiated the intelligibility of putatively synthetic a priori propositions, insisted that the only necessity is logical necessity and denied any sense to the propositions of logic. Hence it seemed to provide the foundations for what the Vienna Circle hailed triumphantly as "consistent empiricism," for it denied that pure reason alone can attain any knowledge of the world. It held metaphysics to be nonsense (the Circle averted their gaze from, or quickly condemned and passed over (Neurath), or attempted to circumvent (Carnap), its paradoxical ineffability claims). And it allocated to philosophy a *sui generis* analytic role and a status wholly distinct from that of science. Schlick, the leading figure in the Circle, went so far as to characterize the *Tractatus* as "the turning point in philosophy," the deepest insight into what the task and status of philosophy should be.

Wittgenstein's influence upon the Vienna Circle was second to none. Indeed, the principle of verification itself was derived from conversations with Wittgenstein in 1929/30, and read back into the *Tractatus*. Members of the Circle spent two academic years reading through the book line by line, abandoning some of its claims and accepting others. They abandoned the picture theory of the proposition, the doctrine of showing and saying, and most of the ontology of logical atomism. But what they accepted was crucial: the account of the nature and limits of philosophy, the conception of logic and logical necessity, and the program of the logical analysis of language (see AYER, CARNAP, HEMPEL, and QUINE). These ideas, interpreted and sometimes seriously misinterpreted, were pivotal to their work. The most important misinterpretation concerned the *Tractatus* account of logic. Members of the Circle agreed with the criticisms of the Fregean and Russellian misconceptions of the nature of logic, and welcomed the view that the propositions of logic are vacuous (senseless). But they gave a conventionalist interpretation to Wittgenstein's account of logic which was far removed from his conception. They thought of the logical connectives as arbitrary symbols introduced to form molecular propositions, whereas Wittgenstein had argued that they are essentially given by the mere idea of an elementary proposition. Where he viewed the truths of logic as flowing from the essential bipolarity of the proposition, they conceived of them as following from the truth-tabular definitions of the logical connectives – hence as true in virtue of the meanings of the logical operators. A logical truth therefore was held to be the logical consequence of conventions (definitions). Wittgenstein, by contrast, had argued that the senseless truths of logic reflect the logical structure of the world. Logic, far from being determined by convention, is transcendental. In the

1930s, when he turned to reconsider his earlier conception, Wittgenstein not only reformulated his views but also vehemently criticized the conventionalism of the Circle. Far from *following* from the meanings of the logical connectives, the truth of the propositions of logic, he argued, is *constitutive* of their meanings.

The collapse of the *Tractatus* vision

Already in the *Tractatus* Wittgenstein had taken note of the fact that determinates of a determinable, e.g. red and green, are mutually exclusive: if A is red all over, it follows that it is not green (or blue or yellow, etc.) all over. At the time, he thought that this showed that "A is red" is not an elementary proposition, and that its entailments would, on analysis, be clarified as following from its truth-functional composition out of elementary propositions. When he returned to philosophy after a hiatus of a decade, he realized that this was misconceived. There are irreducible logical relations of exclusion or implication which are determined not by truth-functional composition, but by the *inner structure* of elementary propositions. He tried to budget for this by abandoning the topic neutrality of the logical connectives and drawing up truth-tables specific to the "propositional system" (i.e. the system of determinates of a determinable) to which a given elementary proposition belongs. In the case of color, the conjunction of "A is red all over" and "A is green all over" is nonsense. Hence the truth-value assignment "TT" must be excluded from such conjunctions by a special rule of syntax. But this concession, he rapidly realized, spells the death-knell for the philosophy of logical atomism, and strikes at the heart of the *Tractatus*. For the independence of the elementary proposition was the pivot upon which turned the whole conception of logic and the ineffable metaphysics of the book. Without it, the idea that the logic of propositions depends only upon the bipolarity of the elementary proposition collapses. The significance of the T/F notation as revealing the essential nature of logical propositions and relations evaporates, precisely because there are logical relations that depend upon the inner structure of elementary propositions. Since the logical operators are not topic neutral, separate truth-tables would have to be drawn up for each propositional system. The idea that there is a general propositional form, according to which every proposition is a result of successive applications to elementary propositions of the operation of joint negation must likewise be relinquished. So too must the thought that generality can be analyzed into logical sums and products, and that the quantifiers can be given a uniform topic-neutral analysis.

As the logical theory of the *Tractatus* collapsed, so too did the metaphysics. It was wrong to say that the world consists of facts rather than of things. Rather, a description of the world consists of statements of facts, not of an enumeration of things. But the statement of a fact just is a true statement. One cannot point *at*, but only point *out*, a fact. And to point out a fact just is to point out that things are thus and so, that is, to make a true assertion. Facts are not concatenations of objects. Unlike concatenations of objects, and unlike states of affairs, facts have no spatiotemporal location. The fact that a circle is red is not composed of redness and circularity concatenated together, since facts are not composed of anything and do not have "constituents." The proposition that *p* is only "made true" by the fact that *p* in the sense in which being a bachelor makes one unmarried. All it means is that the proposition that *p* is true if, in fact,

things are as it says they are. The conception of absolutely simple sempiternal objects was incoherent. For the notions of simplicity and complexity are relative, not absolute. To call spatiotemporal points, properties, or relations "objects" is a misuse of language. What had appeared to be objects that *had* to exist are in fact *samples* which we employ in explaining the meanings of certain ostensively defined expressions in the language. As such, they belong to the means of representation, not (like the postulated "objects" of the *Tractatus*) to what is represented.

As the metaphysics collapsed, so too did the picture theory, the conception of isomorphism between language and reality, and the account of intentionality. What had seemed like an internal relation between the proposition that *p* and the fact that *p* which makes it true was no more than the shadow cast upon reality by an intra-grammatical relation between the expressions "the proposition that *p*" and "the proposition made true by the fact that *p*." There is an internal relation here, but it is forged *in* language – in the grammatical rule that permits the inter-substitution of these expressions – not *between* language and reality. Hence it was mistaken to think that reality *must* have a certain metaphysical form which *must* be reflected in the logico-syntactical forms of language in order for this internal relation to obtain. The intentionality of thought and proposition, which had seemed to demand a pre-established metaphysical harmony between language and reality, is fully explained by reference to intra-grammatical connections between expressions. The thought or expectation that it will be the case that *p* does not "anticipate reality"; rather, only what satisfies the description "it is the case that *p*" will be *called* "the fulfillment of the expectation that it will be the case that *p*." Of course one can "read off" from the thought what will make it true, since the expression of the thought contains the description of the state of affairs the obtaining of which is *called* "the confirmation of the thought." Of course what one thinks, when one thinks that *p*, is what is the case when one's thought is true, but this is not a strange form of identity or coincidence between a shadowy possibility and an actuality. Rather the question "What is being thought?" and "What is the case?" here receive the same answer.

The metaphysics of symbolism of the *Tractatus* was in fact a mythology of symbolism. The meaning of a name is not an object of any kind. What is legitimate about the role which the *Tractatus* simple object was invoked to fulfill is in fact played by defining samples used in ostensive definitions, e.g. of color words. But the sample pointed at in the ostensive definition "This is black" is part of the means of representation, to be used as an object of comparison and standard of correct application of the word "black." Names derive their meanings not from objects in the world which they represent, but from explanations of meaning, of which ostensive definitions are but one type. But it is at best vacuous to claim that all nonlogical terms are names. There are indefinitely many grammatically different kinds of expressions, which fulfill different roles in a language and have different uses, given by the explanations of their meanings, which are in effect rules for their use. In the sense in which the *Tractatus* claimed that there is a connection – a meaning-endowing connection – between language and reality, there is *no* such connection. It was mistaken to suppose that a propositional sign is a fact, that only facts can represent facts, or that only "simple names" can represent simple objects. Far from the logical syntax of language having to mirror the logical forms of things, the different grammars of different languages are autonomous. They owe no homage to reality. They do not reflect language-independent metaphysical possibilities,

determined by the essential nature of objects represented, but rather themselves determine logical possibilities, i.e. what it makes sense to say. Empirical propositions are indeed characteristically (although not uniformly) bipolar, but the concept of a proposition is a *family resemblance* concept: there are many different kinds of proposition, which are not characterized by an essential nature, but by overlapping similarities. The concept of logical form which had informed the *Tractatus* is chimerical. For paraphrase into a canonical notation (as in Russell's theory of descriptions) is not an analysis of what is already present in the paraphrased proposition or thought but a redescription in a different form of representation. Logical form is no reflection of the logico-metaphysical forms of reality, since there is no such thing.

Already in the *Tractatus* Wittgenstein had rejected the logicism in the philosophy of mathematics which Frege and Russell had endeavored unsuccessfully to prove. He denied that numbers were logical objects or reducible to classes. Mathematical propositions, he claimed, are not descriptions of possible states of affairs. Nor are they bipolar. They are, in effect, nonsensical pseudo-propositions; they do not have a sense consisting in their agreement and disagreement with the existence and nonexistence of states of affairs. Rather, they are substitution-rules for the transformation of one empirical proposition concerning magnitudes or quantities or spatial relations, etc. into another, and expressions of rules are not propositions. In the 1930s he wrote extensively about the foundations of mathematics. It is not possible here to do more than indicate briefly the general trajectory of his thought. He did not reject logicism in order to embrace what seemed to be the only alternatives, namely intuitionism and formalism. His fundamental claim is radical. With the liberalization in his concept of a proposition, he was now willing to speak of mathematical propositions. Nevertheless, they are radically unlike empirical propositions, and equally unlike logical ones. Mathematics is a system of interlocked propositions. As already implied in the *Tractatus*, the fundamental role of this system (but not of every proposition within it) is to constitute *rules* for the transformation of empirical propositions. An arithmetic equation, such as $25^2 = 625$, is a rule licensing the transformation of such an empirical proposition as "There are 25 boxes each containing 25 marbles" into the proposition "There are 625 marbles." A theorem of geometry is a norm of representation: a rule permitting the transformation of empirical propositions about shapes, distances, or spatial relations. Different geometries are not different *theories* about empirical space, which might turn out to be true or false. Nor are they different uninterpreted calculi. Rather, they are different *grammars* for the description of spatial relations. Proof *by* mathematics (e.g. in engineering) is wholly different from proof *in* mathematics. While a mathematical proposition is a rule, unless it is an axiom, it is not stipulated, but produced according to rules by a proof. Here we must distinguish proofs within a proof system, e.g. a computation, which is just "homework," as Wittgenstein put it, from proofs which extend mathematics by extending a proof system. Proofs that extend mathematics create new internal relations, modifying existing concepts by linking them with concepts with which they were hitherto unconnected, or connecting them with concepts in new ways – thus licensing novel transformations of appropriate empirical (or other mathematical) propositions. Mathematics is *concept formation*. The propositions of mathematics determine the concepts they invoke. What we conceive of as mathematical necessity is at best a distorted reflection of the inter-

nal relations within a proof system. Mathematics is a human creation, invented rather than discovered.

The *Philosophical Investigations*

Dismantling the *Tractatus* preoccupied Wittgenstein in the early 1930s. Gradually a new method and a wholly different conception of language, of linguistic meaning, and of the relation between language and reality emerged. It became clear that his neglect of questions in the philosophy of psychology in the *Tractatus*, which he had taken to be licensed by the anti-psychologism he took over from Frege, was unwarranted. For the concepts of linguistic meaning are bound up with the concepts of understanding, thinking, intending, and meaning something, and these pivotal notions demand philosophical elucidation. The new method also led to a new conception of philosophy itself – related to, but still importantly different from, the conception of philosophy advocated in the *Tractatus*. That in turn led to a different criticism of metaphysics.

Successive efforts to compose a book laying forth his new ideas culminated in the composition of the *Philosophical Investigations*, Part 1, which was virtually completed by 1945/6. It is his masterwork. Despite some continuities of theme and negative conceptions, it stands in stark contrast not only to the sibylline style of the *Tractatus* but above all to its spirit. Where the *Tractatus* strove for a sublime insight into the language-independent essences of things, the *Investigations* proceeded by a quiet weighing of linguistic facts in order to disentangle knots in our understanding. The *Tractatus* was possessed by a vision of the crystalline purity of the logical forms of thought, language, and the world, the *Investigations* was imbued with a sharpened awareness of the motley of language, the deceptive forms of which lead us into confusion. The *Tractatus* advocated conceptual geology, hoping to disclose the ineffable essences of things by depth analysis of language, the *Investigations* practiced conceptual topography, aiming to dissolve philosophical problems by a patient description of familiar linguistic facts. The *Tractatus* was the culmination of a tradition in western philosophy. The *Investigations* is virtually without precedent in the history of thought.

Wittgenstein's later work, as he himself said, is not merely a stage in the continuous development of philosophy, but constitutes a "kink" in the development of thought comparable to that which occurred when Galileo invented dynamics; it was, in a sense, a new subject, an heir to what used to be called "philosophy." A new method had been discovered, and for the first time it would now be possible for there to be "skillful" philosophers – who would apply the method. The transition from the *Tractatus* to his later philosophy, as he wrote when his new ideas were dawning in 1929, is the transition from the method of truth to the method of meaning. It is a transition from *Wesensschau* – putative insights into the nature or essence of things – to the clarification of conceptual connections in the grammar of our languages, with the purpose of disentangling knots in our thought. The conception of philosophy advocated in the *Investigations* has no precedent, although it is, in a qualified sense, anticipated by the *Tractatus* program for future philosophy. The philosophy of language is equally without ancestors: it is neither a form of idealist telementational linguistic theory (on the model of classical empiricism or de Saussure) nor a form of behaviorist linguistic theory, it is neither a realist truth-conditional semantics nor a form of "anti-realist" semantics. The

philosophy of mind repudiates both dualism as well as mentalism on the one hand and logical behaviorism as well as physicalism on the other. The critique of metaphysics rests neither on Humean or verificationist grounds, nor does it resemble the Kantian critique of transcendent metaphysics. It is no wonder that Wittgenstein's later philosophy has been so frequently misunderstood and misinterpreted, for it can no more be located on received maps of philosophical possibilities than the North Star can be located on a terrestrial globe.

The *Investigations* opens with a quotation from St. Augustine's autobiography in which he recounts the manner in which he assumes that he had learnt to speak. These unselfconscious, nonphilosophical reflections seemed to Wittgenstein to crystallize an important proto-picture of language, a pre-philosophical conception of its role and function, which informs a multitude of philosophical theories. According to this picture the essential role of words is to name things, and the essential role of sentences is to describe how things are. Words are connected to things by means of ostension. This proto-picture, which is akin to an unnoticed field of force unconsciously moulding the shape of sophisticated philosophical theories, is one root of extensive misconceptions in philosophy of language, logic, mathematics, and psychology. It is a muted leitmotif running through the book, and combating the influence of this picture is one of the central tasks of the book. For we are prone to think that corresponding to every name, or corresponding to every name on analysis, there must exist some thing: that nouns name objects, adjectives name properties, verbs name actions, that psychological expressions such as "pain" name psychological objects, and "believe," "want," "intend," "think," etc. name psychological states or processes, number words name numbers, and logical connectives name binary relations. We are inclined to believe that every declarative sentence describes something: that logical propositions describe relations between thoughts, that mathematical propositions describe relations between numbers, that what we conceive of as metaphysical propositions describe necessary relations between ultimate categories of being, that psychological propositions in the first-person describe states of mind, and so on. But this is illusion.

Philosophy of language

The philosophy of language of the *Investigations* has a destructive and a constructive aspect. Its destructive aspect is concerned with undermining the conception of analysis that had informed the *Tractatus* and, more remotely, has characterized philosophy at least since the Cartesian and empiricist programs of analysis into simple natures and into simple ideas respectively. It aims to destroy the conception of a language as a calculus of meaning rules and the idea that the meaning or sense of a sentence is composed of the meanings of its constituent words and derivable from them, given their mode of combination. Hence too, it combats the ideal of determinacy of sense, and the thought that all expressions are either definable by analytic definition or are indefinables and hence explained by an ostensive definition, conceived of as linking language with reality and laying the foundations of language in simple objects given in experience.

It has already been noted that the concepts of simple and complex are relative. Hence whether an A is complex or simple has to be determined by reference to criteria of

simplicity and complexity laid down for As – if there are such criteria. But we commonly confuse the absence of any criteria of complexity (since none have been laid down) with the satisfaction of criteria of simplicity. We are prone to think that an expression is complex if it is defined by analytic definition, and simple if it is explained by ostension. But analytic and ostensive definitions are neither exclusive nor exhaustive. We can explain what "circle" means by saying that *this* is a circle, or by saying that a circle is a locus of points equidistant from a given point. And we can explain what words mean by contextual paraphrase, contrastive paraphrase, exemplification, by a series of examples together with a similarity rider, by gesture, and so on. The meaning of a word is not an object for which a word stands or of which it is the name. Rather, it is what is given by an explanation of meaning, and an explanation of meaning is a rule for the use of the explanandum – a standard of correctness for its application. To ask for the meaning of a word is to ask how it is to be used. Indeed, the meaning of a word is (or, more cautiously, is determined by) its use.

Ostensive definition is one legitimate manner of explaining the meanings of some words. It is not especially privileged: as argued, it does not "connect language with reality" or lay the foundations of language; it is only one rule for the use of the word in question, and it is as capable of being misunderstood as any other explanation of meaning. Many expressions do not have necessary and sufficient conditions of application. Among these are family-resemblance concepts, such as "game," which are explained by a series of examples and a similarity rider. (Even if someone comes up with a sharp definition of "game," that definition is not the rule by reference to which we have been applying the word "game" and by reference to which we would have justified our use of the word.) Indeed, many of the pivotal concepts in philosophy, such as "language," "proposition," "number," "rule," "proof," as well as many psychological concepts, are family-resemblance concepts. Their extension is not determined by common properties, but by overlapping similarities – like the fibers in a rope.

Since numerous kinds of expression are not explained in terms of necessary and sufficient conditions of application, the idea that vagueness is only a surface grammatical feature of language or that it *must* be an imperfection in language is awry. The Fregean demand for determinacy of sense was incoherent. For determinacy of sense is not merely the absence of vagueness, but the exclusion of the very possibility of vagueness: the exclusion, by a complete explanation of meaning, of every possibility of doubt in every conceivable circumstance. But there is no such thing. There is no absolute conception of completeness. The concepts of complete and incomplete are both relative and correlative. A complete explanation of meaning is an explanation which may be invoked as a standard of application in all normal contexts. Relative to that standard, explanations may be judged to be complete or incomplete. But we have no *single* ideal of exactness; what counts as exact or vague varies from context to context. Moreover, vagueness is not always a defect ("I ask him for a bread knife," Wittgenstein mocked, "and he gives me a razor blade because it is sharper"), and its occurrence is not logically "contagious."

The idea that the sense of a sentence is a function of the meanings of its constituents and their mode of composition is a distorted statement of the platitude that if one does not know what the words of a sentence mean or does not understand the way in which they are combined, then one will not understand what is said. The supposition

that what a sentence means follows from an explanation of what its words mean, together with a specification of its structure, errs with regard to both meaning and understanding. The meaning of a sentence is no more composed of the meanings of its parts than a fact is composed of objects. The distinctions between sense and nonsense are not drawn once and for all by reference to circumstance-invariant features of type-sentences, but by reference to circumstance-dependent features of the use of token-sentences. Sentences of precisely the same form may have very different uses. Indeed, the forms of sentences, no matter whether in natural language or translated into a canonical notation of the predicate calculus, conceal rather than reveal their use. Moreover, understanding a sentence is not a process of deriving its meaning from anything.

Little remains of analysis as previously understood. Philosophical problems are misunderstandings caused, among other things, by misleading analogies between forms of expressions with different uses. Some of these can be dissolved by paraphrase, as exemplified by Russell's Theory of Descriptions (see RUSSELL). But it was an illusion that there is anything like a final analysis of the forms of our language, let alone that analysis reveals the logical structure of the world. Instead of analysis as classically conceived, what is needed is a description of the uses of words that will illuminate philosophical confusion, and a rearrangement of familiar rules for the use of words which will make the grammar of the relevant expressions surveyable. For the main source of philosophical puzzlement and of misguided philosophical theories is our failure to command a clear view of the use of words and our consequent entanglement in the network of grammar. Connective analysis (the term is Strawsonian rather than Wittgensteinian), that is, a description of the conceptual connections and exclusions in the web of words, and therapeutic analysis (see below) replace reductive analysis. A sentence is completely analyzed, in the new sense, when its grammar is laid out completely clearly.

A language is misrepresented if it is conceived to be a calculus of rules. More illuminating is the idea that it is a motley of language games. Language is indeed rule-governed, in the loose manner in which games are. Using sentences is comparable to making moves in a game, and a language can be fruitfully viewed as a motley of language games. The use of language is interwoven with the lives and practices of speakers, and is partly constitutive of their form of life. Training and teaching underpin the mastery of a language, and these presuppose shared reactive and behavioral propensities within a linguistic community. Words are like tools, and the diversity of their use is as great as that of different tools: hence masked by conceiving of them as essentially names of things, and concealed by their grammatical form. The greatest error of philosophers of his day, Wittgenstein remarked, is to attend to the forms of expressions rather than to their uses. Even declarative sentences are used for endlessly diverse purposes, of which describing is only one, and non-declarative sentences are misrepresented if taken to be analyzable into a force-indicative component (e.g. an assertion sign or interrogative sign) and a descriptive, truth-value bearing "sentence-radical." Moreover, the concept of description is itself non-uniform, for describing a scene is altogether unlike describing a dream, describing the impression something made is unlike describing the item that made the impression, and describing what one intends is altogether unlike describing the execution of one's

intention. These are logically distinct kinds of descriptions, with distinct kinds of grounds and consequences.

Understanding is akin to an ability, not a state from which performance flows. The criteria for linguistic understanding are of three general kinds: correct use, giving a correct explanation of meaning in context, and responding appropriately to the use of an expression. Viewing explanations of meaning as rules for the use of words, the use of words as rule-following, and understanding as the mastery of the technique of the use of words requires that these concepts be tightly interlocked. And so they are. There is an internal relation between a rule and what counts as compliance with it, which is manifest not only in the interpretations one might give of the rule, but above all in the practice of acting in accordance with it, and in the critical practices of teaching the meanings of expressions, of correcting misapplications and mistaken explanations of meaning. Meaning is determined by use, it is given by an explanation of meaning, and it is what is understood when the meaning of an expression is understood. Not every difference in use is a difference in meaning, but every difference in meaning is a difference in use. Wittgenstein's later philosophy of language is guided by this series of conceptual connections, the ramifications of which he explored in detail.

Philosophy of mind

Against prevailing tradition, Wittgenstein challenged the inner/outer picture of the mind, the conception of the mental as a "world" accessible to its subject by introspection, the conception of introspection as inner perception, the idea that the capacity to say how things are with us "inwardly" is a form of knowledge (let alone a paradigm of self-knowledge), the thought that human behavior is "bare bodily movement," the notion that voluntary action is bodily movement caused by acts of will, the supposition that explanation of human behavior in terms of reasons and motives is causal, and the pervasive influence of the Augustinian picture of language that disposes one to think that psychological expressions are uniformly or even typically names of inner objects, events, processes, or states. His philosophy of mind and of action can be seen as providing a rigorous philosophical underpinning for the hermeneutic insistence on the autonomy of humanistic understanding and its categorial differentiation from understanding in the natural sciences.

Psychological expressions are not names of entities which are directly observable only by the subject, and avowals of the inner are not descriptions of something visible only in a private peepshow. It is all too easy to think of psychological expressions as names of inner entities, and hence of assigning them meaning by private ostensive definition. Wittgenstein's "private language arguments" are aimed at this misconception. There can be no inner, private, analogue of public ostensive definition. Sensations cannot fulfill the role of samples. So a pain cannot serve as a defining sample for the application of the word "pain." Concentrating one's attention upon one's pain is not a kind of pointing. Remembering a sensation presupposes and so cannot explain the meaning of a sensation-name, and the memory of a sensation cannot serve as an object of comparison for the application of a sensation-word. There is no such thing as applying an expression in accordance with a rule which is in principle incommunicable to anyone else. But the idea of defining a sensation word by reference to a

sensation, conceived of as private and intended to function as a defining sample in an ostensive definition would be such a pseudo-rule – for which there could be no criterion of correct application. Whatever seemed to one to be right would be right, and that means that there is no such thing here as right or correct.

Indeed, the very notion of privacy which informs Cartesian and empiricist conceptions of the mental is misconceived. The mental was taken to be private in two senses: privately owned and epistemically private. Pains, for example, were held to be privately owned, i.e. only I can have my pain, another person cannot have my pain but only a qualitatively identical one. And only I can really know that I have a pain, others can only surmise that I do. Both of these claims are misconceived. To *have* a pain is not to own anything, any more than to have a birthday or a train to catch. The distinction between numerical and qualitative identity, which applies to substances, no more applies to pains (or mental images, thoughts or feelings) than it does to colors. If A is red and B is red, then A and B are the same color; so too, if NN has a throbbing headache in his right temple and MM has a throbbing headache in his right temple, then NN and MM have the same headache – neither numerically the same, nor qualitatively the same, but just the same. To think that what differentiates my pain from yours is that I have mine and you have yours is to transform the owner of the pain into a distinguishing property of the pain – which is as absurd as claiming that two chairs cannot have the same color, since the color of *this* chair belongs to *this* chair and the colour of *that* chair belongs to *that* chair.

The conception of epistemic privacy is equally awry. Far from the "inner" being a field of certain empirical knowledge possessed by the subject, which is better known than, and provides the foundations for, other kinds of empirical knowledge, first-person, present-tense psychological utterances are not generally expressions of knowledge at all. "I know I am in pain" is either an emphatic or concessive assertion that I am in pain, or philosophers' nonsense. In such cases, ignorance, doubt, mistake, misidentification, misrecognition are ruled out by *grammar*: we have no use for such forms of words as "I may be in pain, or I may not – I am not sure, I must find out." But we mistake the grammatical exclusion of ignorance, doubt, etc., for the presence of knowledge, certainty, correct identification, and recognition. Whereas they too are excluded as senseless in such cases as pain, and the use of the epistemic operators in other cases has a distinctive meaning; "I don't know what I want" or "I do not know what I believe" are not expressions of ignorance but of indecision. I do not need to look into my mind to find out what I want or believe, but to make it up. If I do not know what I believe about X, I need to examine the evidence, not my state of mind. The utterances "I am in pain," "I'm going to V," "I want G" are standardly employed as *expressions* or *avowals* (rather than descriptions) of pain, intention, or desire, and the utterance is a *criterion* for others to ascribe to the speaker the relevant psychological predicate.

A criterion for the inner is logically (conceptually), as opposed to inductively, good evidence (justification) for ascribing to another an appropriate psychological predicate. Pain and pain behavior, or desire and conative behavior, are not analogically, inductively, or hypothetically connected. Rather, crying out in circumstances of injury, assuaging an injured limb, avoiding the cause of injury, etc. are non-inductive grounds for pain-ascriptions. Grasping the concept of pain involves recognizing such criteria as

grounds for ascription of pain to another. The criteria for ascription of psychological predicates are partly constitutive of the relevant concepts. Psychological utterances or avowals of the inner are (in certain cases) learnt extensions of primitive behavior that manifests the inner. For example, an avowal of pain is grafted onto, and is a partial replacement of, a groan of pain; and while an utterance of pain is as groundless as a shriek of pain, it too constitutes a criterion for third-person ascriptions. It is misguided to suggest that we can never know whether another is in pain. On the contrary, we often know with complete certainty. When someone severely injured screams with pain, just try to doubt whether he really is in pain! Self-knowledge is a hard won achievement, not gained merely by having toothache, wanting or thinking this or that, and being able to say so. Indeed, others often know and understand us better than we do ourselves.

The mind is not a substance. It is not identical with the brain. It is not a private space, in which mental objects are paraded, disclosed to introspective vision. There is, to be sure, such a thing as introspection, but it is not inner perception. Rather it is a form of reflection on one's past, one reasons and motives, affections and attitudes. The third-person pronoun refers neither to the mind nor to the body, but to the person, the living human being. The first-person pronoun functions quite differently; here reference failure, misidentification, misrecognition, and indeterminacy of reference are standardly excluded. "I" is at best a degenerate, limiting case of a referring expression, as a tautology is a limiting case of a proposition with a sense.

Psychological predicates are neither predicable of the body nor of its parts. It is senseless to ascribe to the brain predicates applicable only to the whole creature, e.g. thinking, believing, wanting, or intending. For the criteria for the third-person ascription of such predicates are distinctive forms of *behavior* of the creature in the stream of life, and there is no such thing as a brain manifesting thought or thoughtlessness, belief or incredulity, desire or aversion, intention or inadvertence in what it does. Hence too, it makes no sense to ascribe thought or thoughtlessness, understanding, misunderstanding or failure of understanding to machines. Thought is essentially bound up with the sentient, affective, and conative functions of a being that has a welfare, is capable of desiring and suffering, can set itself goals and pursue them, and can hope to succeed or fear to fail in its projects.

Human behavior that constitutes criteria for the ascription of psychological predicates is not "bare bodily movement," from which we infer analogically or hypothetically their inner state or which we *interpret* as action. On the contrary, we *see* the pain in the face of the sufferer, hear the joy in the voice of a joyful person, perceive the affection in the looks of lovers. Pain, *pace* behaviorists, is not pain-behavior, any more than joy is the same as joyous behavior or love the same as a loving look. But the "inner" is not hidden behind the "outer"; it may sometimes be concealed or suppressed (or it may just not be manifested). But if it is manifested, then it *infuses* the "outer," which is not bare bodily movement, but the actions and affective reactions of living sentient beings in the stream of life. These are not typically describable save in the rich vocabulary of the "inner."

Human action is not movement caused by acts of will. There are such things as acts of will and great efforts of will, but they are unusual, and are not causal antecedents of action. There is such a thing as will power, but that is a matter of tenacity rather

than a psychic analogue of muscle power. Voluntary actions are not actions, let alone movements, preceded by an act of will. Wanting and willing are not names of mental acts or processes, and "He V'ed because he wanted to" does not give a causal explanation of his action; on the contrary, it typically precludes one. Voluntary movement is action for which it makes sense to ask for agential reasons, which a person can decide to perform, try to execute, or be ordered to do. It is marked by lack of agential surprise, and the agent can be held responsible for it.

A reason for action or for belief is a premise in reasoning. Hence it is no more causally related to the action for which it is a reason than the reasons for a belief are causally related to the conclusion which they support. A person's reason is given by specifying the reasoning he went through antecedently to acting or the reasoning he could have gone through and is willing to give *ex post actu*. Reasons, unlike causes, justify or purport to justify that for which they are reasons. A person's avowal of a reason for his action, unlike his typical assertion of a cause of some event, is not a hypothesis. Unlike the assertion of a cause, in the standard case of an avowal of a reason, there is no room for mistake. What makes the connection between the reason and the action is the agent's avowal itself. In avowing a reason, the agent typically takes responsibility for his action viewed under the aspect of the avowed reason.

The critique of metaphysics and nature of philosophy

The *Tractatus* program for future philosophy advocated a non-cognitive conception of philosophy, denying that there could be any philosophical propositions, *a fortiori* any philosophical knowledge. Philosophy should be an activity of elucidation by analysis. Although philosophy was deprived of the possibility of stating essential truths about the natures of things, these very truths were held to be shown by the well-formed propositions of a language, and arriving at a correct logical point of view would include apprehension and appreciation of what cannot be said but shows itself (including truths of ethics and aesthetics).

The later conception of philosophy adhered to the radical non-cognitivism, but rejected the doctrine of linguistically manifest ineffabilia. There are indeed no philosophical truths. What appear as such, and what were construed by the *Tractatus* as an attempt to say what can only be shown, are in effect expressions of rules for the use of expressions in the misleading guise of metaphysical descriptions of the nature of things. So the portentous, apparently metaphysical, claim that the world consists of facts not of things amounts to the grammatical statement that a description of the world consists of a statement of facts and not a list of things. And that in turn is just a rule for the use of the expression "a description of the world." Insofar as metaphysics is conceived to be the quest for knowledge of the necessary forms and structures of the world or of the mind, it is chimerical. All that can be gleaned from these barren fields are grammatical propositions, that is, expressions of rules for the use of words. There are no such things as "necessary facts," and sentences such as "red is a color," or "space is three dimensional" are in effect rules. If something is said to be red, then it can be said to be colored; if something is in space, then its location is given by three coordinates; and so on. Similarly, apparently synthetic a priori truths, such as "Black is darker than white" or "Red is more like orange than it is like yellow," are not insights

into language-independent necessities in the world, but expressions of rules that are partly constitutive of the meanings of the constituent expressions. For any ordered pair of samples which can be used to define "black" and "white" ostensively can also be used to define the relation "darker than." So if *a* is black and *b* is white, it follows without more ado that *a* is darker than *b*. If *a* is red, *b* orange, and *c* yellow, then *a* is more akin to *b* in color than to *c* – one need not look to see. The apparently metaphysical proposition is in fact an inference rule, which is partly constitutive of the meanings of the constituent terms. What appear to be descriptions of objective necessities in the world are merely the shadows cast by the rules for the use of color predicates and relations.

Similarly, "cannot" and "must" in putatively metaphysical propositions mask rules for the use of words. "You cannot travel back in time" or "You cannot count through all the cardinal numbers" look like "An iron nail cannot scratch glass," but they are not. Experience teaches that iron cannot scratch glass. But it is not experience that teaches that one cannot travel in time, rather, it is grammar that stipulates that the form of words "I travelled back to last year" has no use; nothing *counts* as travelling backwards in time. "Cannot" in metaphysics is not about human frailty, but is an expression of a convention. "You cannot count through all the cardinal numbers" is an expression of a grammatical rule which excludes the phrase "counting through all the cardinal numbers" from the language. It does not say that there is something we cannot do, but rather that there is no such thing to do. Similarly, "must" in metaphysics signifies not an objective necessity in reality, but a commitment to a form of representation. "Every event has a cause" is a true or false empirical generalization. "Every event must have a cause" is an expression of a commitment not to call anything "an event" unless it has a (known or unknown) cause.

There are no theories in philosophy, for there can be nothing hypothetico-deductive about the determination of the bounds of sense, nor can it be merely *probable* that such and such a philosophical pronouncement makes no sense. And we do not need to wait upon future confirmation to determine with certainty that it makes no sense. Hence too, there is no philosophical knowledge comparable to knowledge in the sciences. If anyone were to advance theses in philosophy, everyone would agree with them: for example, "Can one step twice into the same river?" – "Yes." Indeed, there are no explanations in philosophy in the sense in which there are in the sciences, for the methods of philosophy are purely descriptive, and not methods of hypothesis formation.

The purified non-cognitivism of the *Investigations* has two aspects. On the one hand, philosophy is a quest for a surveyable representation of a segment of our language with the purpose of solving or dissolving philosophical perplexity. On the other hand, philosophy is a cure for diseases of the understanding. Philosophical problems are conceptual, hence a priori and not empirical. They can be neither solved nor advanced by new information or scientific discoveries, although scientific discoveries may, and often do, raise fresh conceptual puzzles and generate new confusions. Conceptual problems may concern novel concept-formation or existing conceptual structures and relations. The former are exemplified by mathematics, the latter by philosophy. The task of philosophy is to resolve conceptual questions arising out of our existing forms of representation, to clarify conceptual confusions that result from entanglement in the web of the grammar of our language. Philosophy is not a contribution to human

knowledge but to human understanding – an understanding of our forms of representation and their articulations, an overview of the forms of our thought.

The main source of philosophical puzzlement and of misconceived philosophical theories is our failure to command a clear view of the uses of words. The grammar of our language is lacking in surveyability, for expressions with very different uses have similar surface grammars: "I meant" looks akin to "I pointed," "I have a pain" to "I have a pin," "He is thinking" appears akin to "He is talking," "to have a mind" looks like "to have a brain," "2 is greater than 1" seems akin to "Jack is taller than Jill." Hence we misconstrue the meanings of expressions in our philosophical reflections. We think of meaning something or someone as a mental act or activity of attaching signs to objects, take pain to be a kind of object inalienably possessed by the sufferer, imagine that the mind is identical with the brain, assume that statements of numerical inequalities are descriptions, and so on.

What is needed is a perspicuous representation of the segment of grammar that bears on the problem with which we are confronted. It enables us to see differences between concepts that are obscured by the misleadingly similar grammatical forms of expressions. For this no new discoveries are necessary or possible – only the description of grammar, the clarification and arrangement of familiar rules for the use of words. We must *remind* ourselves of what we already know perfectly well, namely how expressions, the use of which we have already mastered, are indeed used. To be sure, these rules must then be arranged in such a manner as to shed light upon the difficulty in question. The rules that concern the philosopher are different from those that concern the grammarian, and the ordering of rules by the philosopher is very different from the ordering sought by the grammarian, for their purposes are quite distinct. A perspicuous representation of a fragment of grammar will enable us to find our way around the relevant part of the grammatical network without stumbling into conceptual confusion. In philosophy, unlike in the sciences, all the information is already at hand – in our knowledge of our language. The problems of philosophy, unlike those of science, are completely solvable. Failure to solve them is due to philosophers' failure to arrange the grammatical facts in such a way that the problems disappear.

Complementary to the conception of philosophy as the quest for a surveyable representation of segments of our language that give rise to conceptual perplexity and confusion is the conception of philosophy as therapeutic. The philosopher's treatment of a question is like the treatment of an illness. One should not try to terminate a disease of thought, either by dogmatism or by the substitution of a technical concept for the problematic one that causes confusion (as Carnap did with his method of "explication"), for slow cure is all important. Every deep philosophical confusion has many different roots, and each must be dug up and examined. Every deeply misconceived answer to a philosophical problem that mesmerizes us and holds us in a vice has many facets, and each must be separately surveyed. Wittgenstein sometimes compared his new methods of philosophical clarification with psychoanalysis. Philosophical theories are latent nonsense; the task of the philosopher is to transform them into patent nonsense. Like the psychoanalyst, the philosopher aims to give the afflicted insight into their own understanding and misunderstanding.

Philosophy is categorially distinct from the sciences. Since there is no philosophical knowledge and there are no licit theories in philosophy, there can be no progress in the

sense in which there is in the sciences. For there is no accumulation of knowledge, no generation of ever richer explanatory theories, no refinement of instrumentation making possible ever more accurate measurement and observation. But there can be progress in another sense, namely in clarification of conceptual structures, in drawing, refining and sharpening distinctions, in destroying conceptual illusions and in eradicating conceptual confusions. However, since there is no way of predicting future forms of entanglement in the web of language, the task of philosophy never ends.

Wittgenstein's place in postwar analytic philosophy

Is Wittgenstein's later philosophy a form of analytic philosophy? The concept of analytic philosophy is neither sharply defined nor uncontested. If one takes the concept of analysis narrowly, connecting it primarily with decompositional analysis, with reduction and logical construction, then one will be inclined to associate analytic philosophy primarily with a variety of forms of philosophy that flourished in the first half of the twentieth century. One will also be prone to associate the movement with a profound interest in, and ingenious philosophical use of, the calculi of formal logic, and, in some cases, in the devising of formal or semi-formal languages to replace the apparently defective natural languages for philosophical purposes. Moore and Russell, the young Wittgenstein of the *Tractatus*, the Cambridge analysts of the early interwar years, and the logical positivists will then be one's paradigmatic analytic philosophers.

Thus construed, it is clear that it would be at best misleading to characterize the later Wittgenstein as an analytic philosopher at all. But it would be perverse to construe analytic philosophy thus. The term "analytic philosophy" was a latecomer upon the scene, and the Oxford philosophers of the postwar era had no qualms in characterizing their work as analytic philosophy and their methods as conceptual analysis. This did not imply that they were dedicated to reductive analysis and logical construction. Indeed, they repudiated them. What it implied was a looser sense of "analysis": the description of the conceptual connections and articulations of salient elements in our conceptual scheme. In this sense, to be sure, the later Wittgenstein *was* an analytic philosopher, and said as much. For, he claimed, a proposition is "fully analyzed" when its grammar has been completely laid bare. Taken in this broader sense, analytic philosophy continued after 1945 in a new and distinctive form. It was dominated by Oxford rather than Cambridge philosophers, although Wittgenstein's philosophy, transmitted to Oxford largely by word of mouth before 1953, was a primary influence upon them (see ANSCOMBE, FOOT, MALCOLM; cf. AUSTIN, RYLE, STRAWSON).

This postwar phase of analytic philosophy lasted for a quarter of a century. It was not a "school" and, unlike the Vienna Circle, issued no manifestos. It was united by its conception of philosophy as an a priori conceptual investigation, contributing to human understanding rather than to human knowledge, hence wholly unlike the sciences. There was consensus that the methodical examination of the use of the relevant words is a *sine qua non* of any serious philosophical investigation. Analytic philosophy of language flourished, as did analytic epistemology; so too did analytic philosophy of psychology and philosophy of action. Paths pioneered by Wittgenstein were followed and refined. But other branches of analytic philosophy, such as analytic jurisprudence, analytic aesthetics, analytic philosophy of history and the social sciences, which had

been of little or no concern to him, were also developed, often in a manner which bore the marks of his influence.

This phase of analytic philosophy waned in the 1970s, and Wittgenstein's influence declined. Whether the forms of philosophy that succeeded it are to be counted as yet another phase of analytic philosophy or as symptoms of its final demise is something that will become clearer only with the passing of time. What is, however, clear, is that Wittgenstein dominated the forms of analytic philosophy from the 1920s until the 1970s, ineradicably impressing his thought upon twentieth-century philosophy.

Notes

1 For example, Alice Ambrose, Elizabeth Anscombe, Max Black, Richard Braithwaite, Karl Britton, Peter Geach, Austin Duncan-Jones, Casimir Lewy, Margaret MacDonald, Norman Malcolm, G. A. Paul, Rush Rhees, Stephen Toulmin, John Wisdom, Georg Henrik von Wright.

2 The other influences upon his thought which he cited retrospectively in 1931 were Karl Kraus, Adolf Loos, Paul Ernst, and Otto Weininger. In later years he made much use of James's *The Principles of Psychology*, which he viewed as a useful source of interesting philosophical confusions – hence not so much an influence upon his own ideas as a stimulus to criticism.

3 To this must be added a large quantity of dictations he gave to Friedrich Waismann for the projected joint work *Logik, Sprache, Philosophie* which was intended as the first volume of the Vienna Circle's series *Schriften zur Wissenschaftliches Weltauffassung*, that volume itself, published in English under the title *Principles of Linguistic Philosophy*, and Waismann's notes of conversations with Wittgenstein published under the title *Wittgenstein and the Vienna Circle*.

4 This contrasts with the Fregean and Russellian conception of the nature of the proposition. Frege held that propositions of natural language may lack a truth-value, although they express a sense. In his logically ideal language, *Begriffsschrift*, every proposition must be bivalent (but not bipolar), i.e. either true or false. Russell held propositions to be bivalent.

Bibliography

Works by Wittgenstein

1953: *Philosophical Investigations*, ed. G. E. M. Anscombe and R. Rhees, trans. G. E. M. Anscombe, Oxford: Blackwell Publishers.

1958: *The Blue and Brown Books*, Oxford: Blackwell Publishers.

1961a: *Notebooks 1914–16*, ed. G. H. von Wright and G. E. M. Anscombe, trans. G. E. M. Anscombe, Oxford: Blackwell Publishers.

1961b: *Tractatus Logico-Philosophicus*, trans. by D. F. Pears and B. F. McGuinness, London: Routledge and Kegan Paul.

1967: *Zettel*, ed. G. E. M. Anscombe and G. H. von Wright, trans. G. E. M. Anscombe, Oxford: Blackwell Publishers.

1969: *On Certainty*, ed. G. E. M. Anscombe and G. H. von Wright, trans. D. Paul and G. E. M. Anscombe, Oxford: Blackwell Publishers.

1970: *Lectures and Conversations on Aesthetics, Psychology, and Religious Beliefs*, compiled from notes by Y. Smythies, R. Rhees, and J. Taylor, ed. C. Barrett, Oxford: Blackwell Publishers.

1971: *ProtoTractatus: An Early Version of Tractatus Logico-Philosophicus*, ed. B. F. McGuinness, T. Nyberg, and G. H. von Wright, trans. D. F. Pears and B. F. McGuinness, London: Routledge and Kegan Paul.

1974: *Philosophical Grammar*, ed. R. Rhees, trans. A. J. P. Kenny, Oxford: Blackwell Publishers.
1975: *Philosophical Remarks*, ed. R. Rhees, trans. R. Hargreaves and R. White, Oxford: Blackwell Publishers.
1976: *Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge 1939*, ed. C. Diamond, Sussex: Harvester Press.
1978: *Remarks on the Foundations of Mathematics*, ed. G. H. von Wright, R. Rhees, and G. E. M. Anscombe, trans. G. E. M. Anscombe, rev. ed., Oxford: Blackwell Publishers.
1979a: *Ludwig Wittgenstein und der Wiener Kreis*, shorthand notes recorded by F. Waismann, ed. B. F. McGuinness, Oxford: Blackwell Publishers, 1967. (English translation, *Wittgenstein and the Vienna Circle*, Oxford: Blackwell Publishers.)
1979b: *Wittgenstein's Lectures, Cambridge 1932–35, from the Notes of Alice Ambrose and Margaret MacDonald*, ed. Alice Ambrose, Oxford: Blackwell Publishers.
1980a: *Culture and Value*, ed. G. H. von Wright in collaboration with H. Nyman, trans. P. Winch, Oxford: Blackwell Publishers.
1980b: *Remarks on the Philosophy of Psychology*, vol. I, ed. G. E. M. Anscombe and G. H. von Wright, trans. G. E. M. Anscombe, Oxford: Blackwell Publishers.
1980c: *Remarks on the Philosophy of Psychology*, vol. II, ed. G. H. von Wright and H. Nyman, trans. C. G. Luckhardt and M. A. E. Aue, Oxford: Blackwell Publishers.
1980d: *Wittgenstein's Lectures, Cambridge 1930–32, from the Notes of John King and Desmond Lee*, ed. Desmond Lee, Oxford: Blackwell Publishers.
1982: *Last Writings on the Philosophy of Psychology*, vol. I, ed. G. H. von Wright and H. Nyman, trans. C. G. Luckhardt and M. A. E. Aue, Oxford: Blackwell Publishers.
1988: *Wittgenstein's Lectures on Philosophical Psychology 1946–47*, notes by P. T. Geach, K. J. Shah, and A. C. Jackson, ed. P. T. Geach, Hemel Hempstead: Harvester Wheatsheaf.
1992: *Last Writings on the Philosophy of Psychology*, vol. II, ed. G. H. von Wright and H. Nyman, trans. C. G. Luckhardt and M. A. E. Aue, Oxford: Blackwell Publishers.
1993: *Ludwig Wittgenstein: Philosophical Occasions 1912–1951*, ed. James Klagge and Alfred Normann, Indianapolis: Hackett.

Works by other authors

Baker, G. P. and Hacker, P. M. S. (1980–96), *An Analytic Commentary on the Philosophical Investigations*, 4 vols, Oxford: Blackwell Publishers; vol. 1, *Wittgenstein: Meaning and Understanding* (Baker and Hacker); vol. 2, *Wittgenstein: Rules, Grammar and Necessity* (Baker and Hacker); vol. 3, *Wittgenstein: Meaning and Mind* (Hacker); and vol. 4, *Wittgenstein: Mind and Will* (Hacker).
Glock, H. J. (1996) *A Wittgenstein Dictionary*, Oxford: Blackwell Publishers.
Hacker, P. M. S. (1986) *Insight and Illusion: Themes in the Philosophy of Wittgenstein*, rev. edn., Oxford: Oxford University Press and Bristol: Thoemmes Press.
Kenny, A. J. P. (1973) *Wittgenstein*, London: Penguin Books.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

6

Rudolf Carnap (1891–1970)

SAHOTRA SARKAR

Rudolf Carnap, pre-eminent member of the Vienna Circle, was one of the most influential figures of twentieth-century analytic philosophy. The Vienna Circle was responsible for promulgating a set of doctrines (initially in the 1920s) which came to be known as logical positivism or logical empiricism. This set of doctrines provides the point of departure for most subsequent developments in the philosophy of science. Consequently Carnap must be regarded as one of the most important philosophers of science of this century. Nevertheless, his most lasting positive contributions were in the philosophy of logic and mathematics and the philosophy of language. Meanwhile, his systematic but ultimately unsuccessful attempt to construct an inductive logic has been equally influential since its failure has convinced most philosophers that such a project must fail.

Carnap was born in 1891 in Ronsdorf, near Bremen, and now incorporated into the city of Wuppertal, in Germany.[1] In early childhood he was educated at home by his mother, Anna Carnap (née Dörpfeld), who had been a schoolteacher. From 1898, he attended the Gymnasium at Barmen, where the family moved after his father's death that year. In school, Carnap's chief interests were in mathematics and Latin. From 1910 to 1914 Carnap studied at the universities of Jena and Freiburg, concentrating first on philosophy and mathematics and, later, on philosophy and physics. Among his teachers in Jena were Bruno Bauch, a prominent neo-Kantian, and Gottlob Frege, a founder of the modern theory of quantification in logic. Bauch impressed upon him the power of Kant's conception that the geometrical structure of space was determined by the form of pure intuition. Though Carnap was impressed by Frege's ongoing philosophical projects, his real (and lasting) influence only came later through a study of his writings (see FREGE). Carnap's formal intellectual work was interrupted between 1914 and 1918 while he did military service during World War I. His political views had already been of a mildly socialist/pacifist nature. The horrors of the war served to make them more explicit and more conscious, and to codify them somewhat more rigorously.

Space

After the war, Carnap returned to Jena to begin research. His contacts with Hans Reichenbach and others pursuing philosophy informed by current science began

during this period. In 1919 he read Whitehead and Russell's *Principia Mathematica* and was deeply influenced by the clarity of thought that could apparently be achieved through symbolization. He began the construction of a putative axiom system for a physical theory of space-time. The physicists – represented by Max Wien, head of the Institute of Physics at the University of Jena – were convinced that the project did not belong in physics. Meanwhile, Bauch was equally certain that it did not belong in philosophy. This incident was instrumental in convincing Carnap of the institutional difficulties faced in Germany of doing interdisciplinary work that bridged the chasm between philosophy and the natural sciences. It also probably helped generate the attitude that later led the logical empiricists to dismiss much of traditional philosophy, especially metaphysics. By this point in his intellectual development (the early 1920s) Carnap was already a committed empiricist who, nevertheless, accepted both the analyticity of logic and mathematics and the Frege–Russell thesis of logicism which required that mathematics be formally constructed and derived from logic.

Faced with this lack of enthusiasm for his original project in Jena, Carnap abandoned it to write a dissertation on the philosophical foundations of geometry, which was subsequently published as *Der Raum* (1922). A fundamentally neo-Kantian work, it included a discussion of "intuitive space," determined by pure intuition, independent of all contingent experience, and distinct from both mathematical (or abstract) space and physical space. However, in contrast to Kant, Carnap restricted what could be grasped by pure intuition to some topological properties of space; metric properties and even the dimensionality of space were regarded as empirical matters. In agreement with Helmholtz and Moritz Schlick (a physicist-turned-philosopher, and founder of the Vienna Circle – see below), the geometry of physical space was also regarded as an empirical matter. Carnap included a discussion of the role of non-Euclidean geometry in Einstein's General Relativity Theory. By distinguishing between intuitive, mathematical, and physical spaces, Carnap attempted to resolve the apparent differences between philosophers, mathematicians, and physicists by assigning the disputing camps to different discursive domains. In retrospect, this move heralded what later became the most salient features of Carnap's philosophical work: tolerance for diverse points of view (so long as they met stringent criteria of clarity and rigor) and an assignment of these viewpoints to different realms, the choice between which is to be resolved not by philosophically substantive (for instance, epistemological) criteria but by pragmatic ones.

The constructionist phase

During the winter of 1921, Carnap read Russell's *Our Knowledge of the External World* (1914). According to Carnap's intellectual autobiography (1963a), this work led him, between 1922 and 1925, to begin the analysis that culminated in *Der logische Aufbau der Welt* (1967), which is usually regarded as Carnap's first major work. The purpose of the *Aufbau* was to construct the everyday world from a phenomenalist basis. This is an epistemological choice (§§54, 58).[2] Carnap distinguished between four domains of objects: autopsychological, physical, heteropsychological, and cultural (§58). The first of these consists of objects of an individual's own psychology; the second of physical entities (Carnap does not distinguish between everyday material objects and the

abstract entities of theoretical physics); the third consists of the objects of some other individual's psychology; and the fourth of cultural objects (*geistige Gegenstände*), which include historical and sociological phenomena.

From Carnap's point of view, "[a]n object . . . is called *epistemically primary* relative to another one . . . if the second one is recognized through the mediation of the first and thus presupposes, for its recognition, the recognition of the first" (§54). Autopsychological objects are epistemically primary relative to the others in this sense. Moreover, physical objects are epistemically primary to heteropsychological ones because the latter can only be recognized through the mediation of the former: an expression on a face, a reading in an instrument, etc. Finally, heteropsychological objects are epistemically primary relative to cultural ones for the same reason.

The main task of the *Aufbau* is construction, which Carnap conceives of as the converse of what he regarded as reduction (which is far from what was then – or is now – conceived of as "reduction" in Anglophone philosophy):

> *an object is 'reducible' to others . . . if all statements about it can be translated into statements which speak only about these other objects.* . . . By *constructing* a concept from other concepts, we shall mean the indication of its "constructional definition" on the basis of other concepts. By a *constructional definition* of the concept *a* on the basis of the concepts *b* and *c*, we mean a rule of translation which gives a general indication how any propositional function in which *a* occurs may be transformed into a coextensive propositional function in which *a* no longer occurs, but only *b* and *c*. If a concept is reducible to others, then it must indeed be possible to construct it from them. (§35)

However, construction and reduction present different formal problems because, except in some degenerate cases (such as explicit definition), the transformations in the two directions may not have any simple explicit relation to each other. The question of reducibility/constructibility is distinct from that of epistemic primacy. In an important innovation in an empiricist context, Carnap argues that both the autopsychological and physical domains can be reduced to each other (in his sense). Thus, at the formal level, either could serve as the basis of the construction. It is epistemic primacy that dictates the choice of the former.

Carnap's task, ultimately, is to set up a constructional system that will allow the construction of the cultural domain from the autopsychological through the two intermediate domains. In the *Aufbau*, there are only informal discussions of how the last two stages of such a construction are to be executed. Only the construction of the physical from the autopsychological is fully treated formally. As the basic units of the constructional system Carnap chose what he calls "elementary experiences" (*Elementarerlebnisse*) (*elex*).[3] These are supposed to be instantaneous cross-sections of the stream of experience – or at least bits of that stream in the smallest perceivable unit of time – that are incapable of further analysis. The only primitive relation that Carnap introduces is "recollection of similarity" (*Rs*). (In the formal development of the system, *Rs* is introduced first and the *elex* are defined as the field of *Rs*.) The asymmetry of *Rs* is eventually exploited by Carnap to introduce temporal ordering.

Since the *elex* are elementary, they cannot be further analyzed to define what would be regarded as constituent qualities of them, such as partial sensations or intensity

components of a sensation. Had the *elex* not been elementary, Carnap could have used "proper analysis" to define such qualities by isolating the individuals into classes on the basis of having a certain (symmetric) relationship with each other. Carnap defines the process of "quasi-analysis" to be formally analogous to proper analysis but only defining "quasi-characteristics" or "quasi-constituents" because the *elex* are unanalyzable.[4] Quasi-analysis based on the relation "part similarity" (*Ps*), itself defined from *Rs*, is the central technique of the *Aufbau*. It is used eventually to define sense classes and, then, the visual sense, visual field places, the spatial order of the visual field, the order of colors and, eventually, sensations. Thus the physical domain is constructed out of the autopsychological. Carnap's accounts of the construction between the other two domains remain promissory sketches.

Carnap was aware that there were unresolved technical problems with his construction of the physical from the autopsychological, though he probably underestimated the seriousness of these problems. The systematic problems are that when a quality is defined as a class selected by quasi-analysis on the basis of a relation: (1) two (different) qualities that happen always to occur together (say, red and hot) will never be separated; and (2) quality classes may emerge in which any two members bear some required relation to each other but there may yet be no relation that holds between all members of the class. Carnap's response to these problems was extra-systematic: in the complicated construction of our world from our *elex*, he hoped that such examples would never or only very rarely arise.[5] Nevertheless, because of these problems, and because the other constructions are not carried out, the attitude of the *Aufbau* is tentative and exploratory: the constructional system is presented as essentially unfinished.[6]

By this point of his intellectual development, Carnap had not only fully endorsed the logicism of the *Principia*, but also the form that Whitehead and Russell had given to logic (that is, the ramified theory of types including the axioms of infinity and reducibility) in that work. However, Poincaré also emerges as a major influence during this period. Carnap did considerable work on the conceptual foundations of physics in the 1920s and some of this work – in particular, his analysis of the relationship between causal determination and the structure of space – shows strong conventionalist attitudes (Carnap 1924; see also Carnap 1923 and 1926).

Viennese positivism

In 1926, at Schlick's invitation, Carnap moved to Vienna to become a *Privatdozent* (instructor) in philosophy at the University of Vienna for the next five years. An early version of the *Aufbau* served as his *Habilitationsschrift*. He was welcomed into the Vienna Circle, a scientific philosophy discussion group organized by (and centered around) Schlick, who had occupied the Chair for Philosophy of the Inductive Sciences since 1922. In the meetings of the Vienna Circle the typescript of the *Aufbau* was read and discussed. What Carnap seems to have found most congenial in the Circle – besides its members' concern for science and competence in modern logic – was their rejection of traditional metaphysics. Over the years, besides Carnap and Schlick, the Circle included Herbert Feigl, Kurt Gödel, Hans Hahn, Karl Menger, Otto Neurath, and Friedrich Waismann, though Gödel would later claim that he had little sympathy for the

anti-metaphysical position of the other members. The meetings of the Circle were characterized by open, intensely critical, discussion with no tolerance for ambiguity of formulation or lack of rigor in demonstration. The members of the Circle believed that philosophy was a collective enterprise in which progress could be made.

These attitudes, even more than any canonical set of positions, characterized the philosophical movement, initially known as logical positivism and, later, as logical empiricism, that emerged from the work of the members of the Circle and a few others, especially Reichenbach. However, besides rejecting traditional metaphysics, most members of the Circle accepted logicism and a sharp distinction between analytic and synthetic truths. The analytic was identified with the a priori; the synthetic with the a posteriori. A. J. Ayer, who attended some meetings of the Circle in 1933 (after Carnap had left – see below) returned to London and published *Language, Truth and Logic* in 1936 (see AYER). This short book did much to popularize the views of the Vienna Circle among Anglophone philosophers though it lacks the sophistication that is found in the writings of the members of the Circle, particularly Carnap.

Under Neurath's influence, during his Vienna years, Carnap abandoned the phenomenalist language he had preferred in the *Aufbau* and came to accept physicalism. The epistemically privileged language is one in which sentences reporting empirical knowledge of the world ("protocol sentences") employ terms referring to material bodies and their observable properties. From Carnap's point of view, the chief advantage of a physicalist language is its intersubjectivity. Physicalism, moreover, came hand-in-hand with the thesis of the "unity of science," that is, that the different empirical sciences (including the social sciences) were merely different branches of a single unified science. To defend this thesis, it had to be demonstrated that psychology could be based on a physicalist language. In an important paper only published somewhat later, Carnap (1934b) attempted that demonstration. Carnap's adoption of physicalism was final; he never went back to a phenomenalist language. However, what he meant by "physicalism" underwent radical transformations over the years. By the end of his life, it meant no more than the adoption of a non-solipsistic language, that is, one in which intersubjective communication is possible (Carnap 1963b).

In the Vienna Circle, Wittgenstein's *Tractatus* was discussed in detail. Carnap found Wittgenstein's rejection of metaphysics concordant with the views he had developed independently. Partly because of Wittgenstein's influence on some members of the Circle (though not Carnap), the rejection of metaphysics took the form of an assertion that the sentences of metaphysics are meaningless in the sense of being devoid of cognitive content. Moreover, the decision whether a sentence is meaningful was to be made on the basis of the principle of verifiability, which claims that the meaning of a sentence is given by the conditions of its (potential) verification. Observation terms are directly meaningful on this account. Theoretical terms only acquire meaning through explicit definition from observation terms. Carnap's major innovation in these discussions within the Circle was to suggest that even the thesis of realism – asserting the "reality" of the external world – is also meaningless, a position not shared by Schlick, Neurath, or Reichenbach. Problems generated by meaningless questions became the celebrated "pseudo-problems" of philosophy (Carnap 1967).

Wittgenstein's principle of verifiability posed fairly obvious problems in any scientific context. No universal generalization can ever be verified. Perhaps independently,

Karl Popper perceived the same problem (see POPPER). This led him to replace the requirement of verifiability with that of falsifiability, though only as a criterion to demarcate science from metaphysics, and not as one also to be used to demarcate meaningful from meaningless claims. It is also unclear what the status of the principle itself is, that is, whether it is meaningful by its own criterion of meaningfulness. Carnap, as well as other members of the Vienna Circle including Hahn and Neurath, realized that a weaker criterion of meaningfulness was necessary. Thus began the program of the "liberalization of empiricism." There was no unanimity within the Vienna Circle on this point. The differences between the members are sometimes described as those between a conservative "right" wing, led by Schlick and Waismann, which rejected the liberalization of empiricism, and the epistemological anti-foundationalism that is involved in the move to physicalism; and a radical "left" wing, led by Neurath and Carnap, which endorsed the opposite views. The "left" wing also emphasized fallibilism and pragmatics; Carnap went far enough along this line to suggest that empiricism itself was a proposal to be accepted on pragmatic grounds. This difference also reflected political attitudes insofar as Neurath, and to a lesser extent, Carnap viewed science as a tool for social reform.

The precise formulation of what came to be called the criterion of cognitive significance took three decades. (See Hempel 1950 and Carnap 1956 and 1961.) In an important pair of papers, "Testability and Meaning," Carnap (1936, 1937a) replaced the requirement of verification with that of confirmation; at this stage, he made no attempt to quantify the latter. Individual terms replace sentences as the units of meaning. Universal generalizations are no longer problematic; though they cannot be conclusively verified, they can yet be confirmed. Moreover, in "Testability and Meaning," theoretical terms no longer require explicit definition from observational ones in order to acquire meaning; the connection between the two may be indirect through a system of implicit definitions. Carnap also provides an important pioneering discussion of disposition predicates.

The syntactic phase

Meanwhile, in 1931, Carnap had moved to Prague, where he held the Chair for Natural Philosophy at the German University until 1935 when, under the shadow of Hitler, he emigrated to the United States. Towards the end of his Vienna years, a subtle but important shift in Carnap's philosophical interests had taken place. This shift was from a predominant concern for the foundations of physics to that for the foundations of mathematics and logic, even though he remained emphatic that the latter were important only insofar as they were used in the empirical sciences, especially physics.

In Vienna and before, following Frege and Russell, Carnap espoused logicism in its conventional sense, that is, as the doctrine that held that the concepts of mathematics were definable from those of logic and the theorems of mathematics were derivable from the principles of logic. In the aftermath of Gödel's (1931) incompleteness theorems (see TARSKI, CHURCH, GÖDEL), however, Carnap abandoned this type of logicism and opted, instead, for the requirement that the concepts of mathematics and logic always have their customary, that is, everyday interpretation in all contexts. He also began to advocated a strong conventionalism regarding what constituted "logic."

Besides the philosophical significance of Gödel's results, what impressed Carnap most about that work was Gödel's arithmetization of syntax. Downplaying the distinction between an object language and its metalanguage, Carnap interpreted this procedure as enabling the representation of the syntax of a language within the language itself. At this point Carnap had not yet accepted the possibility of semantics even though he was aware of some of Tarski's work and had had some contact with the Polish school of logic. In this context, the representation of the syntax of a language within itself suggested to Carnap that all properties of a language could be studied within itself through a study of syntax.

These positions were codified in Carnap's major work from this period, *The Logical Syntax of Language* (Carnap 1937b). The English translation includes material that had to be omitted from the German original owing to a shortage of paper; the omitted material was separately published in German as papers (Carnap 1934a, 1935). Conventionalism about logic was incorporated into the well-known Principle of Tolerance:

> *It is not our business to set up prohibitions but to arrive at conventions* [about what constitutes a logic]. . . . *In logic, there are no morals*. Every one is at liberty to build up his own logic, i.e., his own form of language, as he wishes. All that is required is that, if he wishes to discuss it, he must state his method clearly, and give syntactic rules instead of philosophical arguments. (1937b: 51–2; emphasis in the original)

Logic, therefore, is nothing but the syntax of language.

In *Syntax*, the Principle of Tolerance allows Carnap to navigate the ongoing disputes between logicism, formalism, and intuitionism/constructivism in the foundations of mathematics without abandoning any insight of interest from these schools. Carnap begins with a detailed study of the construction of two languages, I and II. The last few sections of *Syntax* also present a few results regarding the syntax of any language and also discuss the philosophical ramifications of the syntactic point of view.[7]

Language I, which Carnap calls "definite," is intended as a neutral core of all logically interesting languages, neutral enough to satisfy the strictures of almost any intuitionist or constructivist. It permits the definition of primitive recursive arithmetic and has bounded quantification (for all x up to some upper bound) but not much more. Its syntax is fully constructed formally. Language II, which is "indefinite" for Carnap, is richer. It includes Language I and has sufficient resources for the formulation of all of classical mathematics and is, therefore, non-constructive. Moreover, Carnap permits descriptive predicates in each language. Thus, the resources of Language II are strong enough to permit, in principle, the formulation of classical physics. The important point is that, because of the Principle of Tolerance, the choice between Languages I and II or, for that matter, any other syntactically specified language, is not based on factual considerations. If one wants to use mathematics to study physics in the customary way, Language II is preferable since, as yet, non-constructive mathematics remains necessary for physics. But the adoption of Language II, dictated by the pragmatic concern for doing physics, does not make Language I incorrect. This was Carnap's response to the foundational disputes of mathematics: by tolerance they are defined out of existence.

The price paid if one adopts the Principle of Tolerance is a radical conventionalism about what constitutes logic. Conventionalism, already apparent in Carnap's admission of both a phenomenalist and a physicalist possible basis for construction in the *Aufbau*, and strongly present in the works on the foundations of physics in the 1920s, had now been extended in *Syntax* to logic. As a consequence, what might be considered to be the most important question in any mathematical or empirical context – the choice of language – became pragmatic. This trend of relegating troublesome questions to the realm of pragmatics almost by fiat, thereby excusing them from systematic philosophical exploration, became increasingly prevalent in Carnap's views as the years went on.

Syntax contained four technical innovations in logic that are of significance: (1) a definition of analyticity that, as was later shown by S. C. Kleene, mimicked Tarski's definition of truth for a formalized language; (2) Carnap constructed a proof, independently of Tarski, that truth cannot be defined as a syntactic predicate in any consistent formalized language; (3) a rule for infinite induction (in Language I) that later came to be called the omega rule; and (4), most important, a generalization of Gödel's first incompleteness theorem that has come to be called the fixed-point lemma. With respect to (4), what Carnap proved is that, in a language strong enough to permit arithmetization, for any syntactic predicate, one can construct a sentence that would be interpreted as saying that it satisfies that predicate. If the chosen predicate is unprovability, one gets Gödel's result.

Besides the Principle of Tolerance, the main philosophical contribution of *Syntax* was the thesis that philosophy consisted of the study of logical syntax. Giving a new twist to the Vienna Circle's claim that metaphysical claims were meaningless, Carnap argues and tries to show by example that sentences making metaphysical claims are all syntactically ill-formed. Moreover, since the arithmetization procedure shows that all the syntactic rules of a language can be formulated within the language, even the rules that determine what sentences are meaningless can be constructed within the language. All that is left for philosophy is a study of the logic of science. But, as Carnap puts it: "The *logic of science* (logical methodology) is nothing else than the *syntax of the language of science*. . . . To share this view is to *substitute logical syntax for philosophy*" (1937b: 7–8). The claims of *Syntax* are far more grandiose – and more flamboyant – than anything in the *Aufbau*.

Semantics

In the late 1930s Carnap abandoned the narrow syntacticism of *Syntax* and, under the influence of Tarski and the Polish school of logic, came to accept semantics. With this move, Carnap's work enters its final mature phase. For the first time, he accepted that the concept of truth can be given more than pragmatic content. Thereupon, he turned to the systematization of semantics with characteristic vigor, especially after his immigration to the US where he taught at the University of Chicago from 1936 to 1952. In his contribution to the *International Encyclopedia of Unified Science*, in 1939, on the foundations of logic and mathematics, the distinctions between syntactic, semantic, and pragmatic considerations regarding any language are first presented in their mature form.

Introduction to Semantics, which followed in 1942, develops semantics systematically. In *Syntax* Carnap had distinguished between two types of transformations on sentences: those involving "the method of derivation" or "*d*-method"; and those involving the "method of consequence" or "*c*-method." Both of these were supposed to be syntactic but there is a critical distinction between them. The former allows only a finite number of elementary steps. The latter places no such restriction and is, therefore, more "indefinite." Terms defined using the *d*-method ("*d*-terms") include "derivable," "demonstrable," "refutable," "resoluble," and "irresoluble"; the corresponding "*c*-terms" are "consequence," "analytic," "contradictory," "L-determinate," and "synthetic." After the conversion to semantics Carnap proposed that the *c*-method essentially captured what semantics allowed; the *c*-terms referred to semantic concepts.

Thus semantics involves a kind of formalization, though one that is dependent on stronger inference rules than the syntactical ones. In this sense, as Church (1956: 65) has perceptively pointed out, Carnap (and Tarski) reduce semantics to formal rules, that is, syntax. Thus emerges the interpretation of deductive logic that has since become the textbook version, so commonly accepted that is has become unnecessary to refer to Carnap when one uses it. For Carnap, the semantic move had an important philosophical consequence: philosophy was no longer to be replaced just by the syntax of the language of science; rather, it was to be replaced by the syntax and the semantics of the language of science.

Carnap's most original – and influential – work in semantics is *Meaning and Necessity* (1947), where the basis for an intensional semantics was laid down. Largely following Frege, intensional concepts are distinguished from extensional ones. Semantical rules are introduced and the analytic/synthetic distinction is clarified by requiring that any definition of analyticity must satisfy the (meta-)criterion that analytic sentences follow from the semantical rules alone. By now Carnap had fully accepted that semantic concepts and methods are more fundamental than syntactic ones: the retreat from the flamboyance of *Syntax* was complete. The most important contribution of *Meaning and Necessity* was the reintroduction into logic, in the new intensional framework, of modal concepts that had been ignored since the pioneering work of Lewis (1918). In the concluding chapter of his book Carnap introduced an operator for necessity, gave semantic rules for its use, and showed how other modal concepts such as possibility, impossibility, necessary implication, and necessary equivalence can be defined from this basis.

By this point, Carnap had begun to restrict his analyses to exactly constructed languages, implicitly abandoning even a distant hope that they would have any direct bearing on natural languages. The problem with the latter is that their ambiguities made them unsuited for the analysis of science which, ultimately, remained the motivation of all of Carnap's work. Nevertheless, Carnap's distinction between the analytic and the synthetic came under considerable criticism from many, including Quine (1951), primarily on the basis of considerations about natural languages. Though philosophical fashion has largely followed Quine on this point, at least until recently, Carnap was never overly impressed by this criticism (Stein 1992). The analytic/synthetic distinction continued to be fundamental to his views and, in a rejoinder to Quine, Carnap argued that nothing prevented empirical linguistics from exploring intensions and thereby discovering cases of synonymy and analyticity (Carnap 1955).

Carnap's most systematic exposition of his final views on ontology is also from this period (1950a). A clear distinction is maintained between questions that are internal to a linguistic framework and questions that are external to it. The choice of a linguistic framework is to be based not on cognitive but on pragmatic considerations. The external question of "realism," which ostensibly refers to the "reality" of entities of a framework in some sense independent of it, rather than to their "reality" within it after the framework has been accepted, is rejected as non-cognitive. This appears to be an anti-"realist" position but is not in the sense that, within a framework, Carnap is tolerant of the abstract entities that bother nominalists. The interesting question becomes the pragmatic one, that is, what frameworks are fruitful in which contexts, and Carnap's attitude towards the investigation of various alternative frameworks remains characteristically and consistently tolerant.

Carnap continued to explore questions about the nature of theoretical concepts and to search for a criterion of cognitive significance, preoccupations of the logical empiricists that date back to the Vienna Circle. In 1956 he published a detailed exposition of his final views regarding the relation between the theoretical and observational parts of a scientific language (Carnap 1956). This paper emphasizes the methodological and pragmatic aspects of theoretical concepts.

It also contains his most subtle, though not his last, attempt to explicate the notion of the cognitive significance of a term and thus establish clearly the boundary between scientific and nonscientific discourse. However, the criterion he formulates makes theoretical terms significant only with respect to a class of terms, a theoretical language, an observation language, correspondence rules between them, and a theory. Relativization to a theory is critical to avoiding the problems that beset earlier attempts to find such a criterion. Carnap proves several theorems that are designed to show that the criterion does capture the distinction between scientific and nonscientific discourse. This criterion was criticized by Roozeboom (1960) and Kaplan (1975) but these criticisms depend on modifying Carnap's original proposal in important ways. According to Kaplan, Carnap accepted his criticism though there is apparently no independent confirmation of that fact. However, Carnap (1961) did turn to a different formalism (Hilbert's ε-operator) in his last attempt to formulate such a criterion and this may indicate dissatisfaction with the 1956 attempt. If so, it remains unclear why: that attempt did manage to avoid the technical problems associated with the earlier attempts of the logical empiricists.

Inductive logic

From 1941 onwards Carnap also began a systematic attempt to analyze the concepts of probability and to formulate an adequate inductive logic (a logic of confirmation), a project that would occupy him for the rest of his life. Carnap viewed this work as an extension of the semantical methods that he had been developing for the last decade. This underscores an interesting pattern in Carnap's intellectual development. Until the late 1930s Carnap only viewed syntactic categories as non-pragmatically specifiable; questions of truth and confirmation were viewed as pragmatic. His conversion to semantics saw the recovery of truth from the pragmatic to the semantic realm. Now, confirmation followed truth down the same pathway.

In *Logical Foundations of Probability* (1950b), his first systematic analysis of probability, Carnap distinguished between two concepts of probability: "statistical probability," which was the relevant concept to be used in empirical contexts and generally estimated from the relative frequencies of events, and "logical probability," which was to be used in contexts such as the confirmation of scientific hypotheses by empirical data. Though the latter concept, usually called the "logical interpretation" of probability went back to Keynes (1921), Carnap provides its first systematic explication.

Logical probability is explicated from three different points of view (1950b: 164–8): (1) as a conditional probability $c(h,e)$ which measures the degree of confirmation of a hypothesis h on the basis of evidence e (if $c(h,e) = r$, then r is determined by logical relations between h and e); (2) as a rational degree of belief or fair betting quotient (if $c(h,e) = r$, then r is a fair bet on h if e correctly describes the total knowledge available to a bettor); and (3) as the limit of relative frequencies in some cases. According to Carnap, the first of these, which specifies a confirmation function ("c-function"), is the concept that is most relevant to the problem of induction. In the formal development of the theory, probabilities are associated with sentences of a formalized language.

In *Foundations*, Carnap believed that a unique measure $c(h,e)$ of the degree of confirmation can be found and he even proposed one (namely, Laplace's rule of succession) though he could not prove its uniqueness satisfactorily. His general strategy was to augment the standard axioms of the probability calculus by a set of "conventions on adequacy" (1950b: 285), which turned out to be equivalent to assumptions about the rationality of degrees of belief that had independently been proposed by both Ramsey and de Finetti (Shimony 1992). In a later work, *The Continuum of Inductive Methods* (1952), using the conventions on adequacy and some plausible symmetry principles, Carnap managed to show that all acceptable c-functions could be parameterized by a single parameter, a real number, $\lambda \in [0,\infty]$. The trouble remained that there is no intuitively appealing a priori strategy to restrict λ to some preferably very small subset of $[0,\infty]$. At one point, Carnap even speculated that it would have to be fixed empirically. Unfortunately, some higher-order induction would then be required to justify the procedure for its estimation and, potentially, this leads to infinite regress.

Carnap spent 1952–4 at the Institute for Advanced Study at Princeton where he continued to work on inductive logic, often in collaboration with John Kemeny. He also returned to the foundations of physics, apparently motivated by a desire to trace and explicate the relations between the physical concept of entropy and an abstract concept of entropy appropriate for inductive logic. His discussion with physicists proved to be disappointing and he did not publish his results.[8]

In 1954 Carnap moved to the University of California at Los Angeles to assume the chair that had become vacant with Reichenbach's death in 1953. There he continued to work primarily on inductive logic, often with several collaborators, over the next decade. There were significant modifications of his earlier attempts to formulate a systematic inductive logic.[9] Obviously impressed by the earlier work of Ramsey and de Finetti, Carnap (1971b) returned to the second of his three 1950 explications of logical probability and emphasized the use of inductive logic in decision problems.

More importantly, Carnap, in "A Basic System of Inductive Logic" (1971a, 1980) finally recognized that attributing probabilities to sentences was too restrictive. If a

conceptual system uses real numbers and real-valued functions, no language can express all possible cases using only sentences or classes of sentences. Because of this, he now began to attribute probabilities to events or propositions (which are taken to be synonymous). This finally brought some concordance between his formal methods and those of mathematical statisticians interested in epistemological questions. Propositions are identified with sets of models; however, the fields of the sets are defined using the atomic propositions of a formalized language. Thus, though probabilities are defined as measures of sets, they still remain relativized to a particular formalized language. Because of this, and because the languages considered remain relatively simple (mostly monadic predicate languages) much of this work remains similar to the earlier attempts.

By this point Carnap had abandoned the hope of finding a unique *c*-function. Instead, he distinguished between subjective and objective approaches in inductive logic. The former emphasizes individual freedom in the choice of necessary conventions; the latter emphasizes the existence of limitations. Though Carnap characteristically claimed to keep an open mind about these two approaches, his emphasis was on finding rational a priori principles which would systematically limit the choice of *c*-functions. Carnap was still working on this project when he died on 14 September 1970. He had not finished revising the last sections of the second part of the "Basic System," both parts of which were only published posthumously.

Towards the end of his life, Carnap's concern for political and social justice had led him to become an active supporter of an African-American civil rights organization in Los Angeles. According to Stegmüller (1972: lxvi), the "last photograph we have of Carnap shows him in the office of this organization, in conversation with various members. He was the only white in the discussion group."

The legacy

Some decades after Carnap's death it is easier to assess Carnap's legacy, and that of logical empiricism, than it was in the 1960s and 1970s when a new generation of analytic philosophers and philosophers of science apparently felt that they had to reject that work altogether in order to be able to define their own philosophical agendas. This reaction can itself be taken as evidence of Carnap's seminal influence but, nevertheless, it is fair to say that Carnap and logical empiricism fell into a period of neglect in the 1970s from which it only began to emerge in the late 1980s and early 1990s. Meanwhile it became commonplace among philosophers to assume that Carnap's projects had failed.

Diagnoses of this failure have varied. For some it was a result of the logical empiricists' alleged inability to produce a technically acceptable criterion for cognitive significance. For others, it was because of Quine's dicta against the concept of analyticity and the analytic/synthetic distinction. Some took Popper's work to have superseded that of Carnap and the logical empiricists. Many viewed Kuhn's seminal work on scientific change to have shown that the project of inductive logic was misplaced (see KUHN); they, and others, generally regarded Carnap's attempt to explicate inductive logic to have been a failure. Finally, a new school of "scientific realists" attempted to escape Carnap's arguments against external realism.

There can be little doubt that Carnap's project of founding inductive logic has faltered. He never claimed that he had gone beyond preliminary explorations of possibilities and though there has been some work since, by and large, epistemologists of science have abandoned that project in favor of less restrictive formalisms, for instance, those associated with Bayesian or Fisherian statistics. But, with respect to every other case mentioned in the last paragraph, the situation is far less clear. It has already been noted that Carnap's final criterion for cognitive significance does not suffer from any technical difficulty no matter what its other demerits may be. Quine's dicta against analyticity no longer appear as persuasive as they once did (Stein 1992); Quine's preference for using natural – rather than formalized – language in the analysis of science has proved to be counterproductive; and his program of naturalizing epistemology is yet to live up to any initial promise that it ever might have had. Putnam's "internal realism" is based on and revives Carnap's views on ontology and Kuhn is perhaps now better regarded as having contributed significantly to the sociology rather than the epistemology of science.

However, to note that some of the traditionally fashionable objections to Carnap and logical empiricism cannot be sustained does not show that that work deserves a positive assessment on its own. We are still left with the question: what, exactly, did Carnap contribute? The answer turns out to be surprisingly straightforward: the textbook picture of deductive logic that we have today is the one that Carnap produced in the early 1940s after he came to acknowledge the possibility of semantics. The fixed-point lemma has turned out to be an important minor contribution to logic. The reintroduction of modal logic into philosophy opened up new vistas for Kripke and others in the 1950s and 1960s (see KRIPKE). Carnap's views on ontology continue to influence philosophers today. Moreover, even though the project of inductive logic seems unsalvageable to most philosophers it is hard to deny that Carnap managed to clarify significantly the ways in which concepts of probability must be deployed in the empirical sciences and why the problem of inductive logic is so difficult. But, most of all, Carnap took philosophy to a new level of rigor and clarity, accompanied by an open-mindedness (codified in the Principle of Tolerance) that, unfortunately, has not been widely shared in analytic philosophy.[10]

Notes

1 Biographical details are from Carnap 1963a.
2 References to the *Aufbau* are to sections; this permits the simultaneous use of the German and English editions.
3 An excellent discussion of Carnap's construction is to be found in Goodman 1951, ch. 5.
4 Thus, if an *elex* is both *c* in color and *t* in temperature, *c* or *t* can be defined as classes of every *elex* having *c* or *t* respectively. However, to say that *c* or *t* is a quality would imply that an *elex* is analyzable into simpler constituents. Quasi-analysis proceeds formally in this way (as if it is proper analysis) but only defines quasi-characteristics thus leaving each *elex* unanalyzable.
5 Goodman (1951) also provides a very lucid discussion of these problems.
6 Some recent scholarship has questioned whether Carnap had any traditional epistemological concerns in the *Aufbau*. In particular, Friedman (e.g. 1992) has championed the view that Carnap's concerns in that work are purely ontological: the *Aufbau* is not concerned with the

question of the source or status of our knowledge of the external world; rather, it investigates the bases on which such a world may be constructed. (See, also, Richardson 1998. Both Friedman and Richardson – as well as Sauer (1985) and Haack (1977) long before them – emphasize the Kantian roots of the *Aufbau*.) If this reinterpretation is correct, then what exactly the *Aufbau* owes to Russell (and traditional empiricism) becomes uncertain. However, as Putnam (private communication) has pointed out, this reinterpretation goes too far: though the project of the *Aufbau* is not identical to that of Russell's external world program (for reasons including those that Friedman gives), there is sufficient congruence between the two projects for Carnap to have correctly believed that he was carrying out Russell's program. In particular, the formal constructions of the *Aufbau* are a necessary prerequisite for the development of the epistemology that Russell had in mind: one must be able to construct the world formally from a phenomenalist basis before one can suggest that this construction shows that the phenomena are the source of our knowledge of the world. Moreover, this reinterpretation ignores the epistemological remarks scattered throughout the *Aufbau* itself, including Carnap's concern for the epistemic primacy of the basis he begins with.

7 Sarkar (1992) attempts a comprehensible reconstruction of the notoriously difficult formalism of *Syntax*.

8 These were edited and published by Abner Shimony (as *Two Essays on Entropy*, 1977) after Carnap's death.

9 See Carnap and Jeffrey 1971 and Jeffrey 1980. An excellent introduction to this part of Carnap's work on inductive logic is Hilpinen 1975.

10 For comments on earlier versions of this essay thanks are due to Justin Garson, Cory Juhl, Al Martinich, and Itai Sher.

Bibliography

Works by Carnap

1922: *Der Raum*, Berlin: von Reuther und Reichard.

1923: "Über die Aufgabe der Physik und die Anwendung des Grundsatzes der Einfachsteit," *Kant-Studien* 28, pp. 90–107.

1924: "Dreidimensionalität des Raumes und Kausalität: Eine Untersuchung über den logischen Zusammenhang zweier Fiktionen," *Annalen der Philosophie und philosophischen Kritik* 4, pp. 105–30.

1926: *Physikalische Begriffsbildung*, Karlsruhe: Braun.

1934a: "Die Antinomien und die Unvollständigkeit der Mathematik," *Monatshefte für Mathematik und Physik* 41, pp. 42–8.

1934b: *The Unity of Science*, London: Kegan Paul, Trench, Trubner and Co. (Originally published 1932.)

1935: "Ein Gültigskriterium für die Sätze der klassischen Mathematik," *Monatshefte für Mathematik und Physik* 42, pp. 163–90.

1936: "Testability and Meaning," *Philosophy of Science* 3, pp. 419–71.

1937a: "Testability and Meaning," *Philosophy of Science* 4, pp. 1–40.

1937b: *The Logical Syntax of Language*, London: Kegan Paul, Trench, Trubner and Co. (Originally *Logische Syntax der Sprache*, Vienna: Springer, 1934.)

1939: *Foundations of Logic and Mathematics*, Chicago: University of Chicago Press.

1947: *Meaning and Necessity: A Study in Semantics and Modal Logic*, Chicago: University of Chicago Press.

1950a: "Empiricism, Semantics, and Ontology," *Revue Internationale de Philosophie* 4, pp. 20–40.

1950b: *Logical Foundations of Probability*, Chicago: University of Chicago Press.
1952: *The Continuum of Inductive Methods*, Chicago: University of Chicago Press.
1955: "Meaning and Synonymy in Natural Languages," *Philosophical Studies* 7, pp. 33–47.
1956: "The Methodological Character of Theoretical Concepts," in *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, ed. H. Feigl and M. Scriven, Minneapolis: University of Minnesota Press, pp. 38–76.
1961: "On the Use of Hilbert's ε-Operator in Scientific Theories," in *Essays on the Foundations of Mathematics*, ed. Y. Bar-Hillel, E. I. J. Poznanski, M. O. Rabin, and A. Robinson, Jerusalem: Magnes Press, pp. 156–64.
1963a: "Intellectual Autobiography," in *The Philosophy of Rudolf Carnap*, ed. P. A. Schilpp, La Salle, IL: Open Court, pp. 3–84.
1963b: "Replies and Systematic Expositions," in *The Philosophy of Rudolf Carnap*, ed. P. A. Schilpp, La Salle, IL: Open Court, pp. 859–1013.
1967: *The Logical Structure of the World and Pseudoproblems in Philosophy*, Berkeley: University of California Press. (Originally published 1928.)
1971a: "A Basic System of Inductive Logic, Part I," in *Studies in Inductive Logic and Probability*, ed. R. Carnap and R. C. Jeffrey, vol. 1, Berkeley: University of California Press, pp. 33–165.
1971b: "Inductive Logic and Rational Decisions," in *Studies in Inductive Logic and Probability*, ed. R. Carnap and R. C. Jeffrey, vol. 1, Berkeley: University of California Press, pp. 5–31.
1977: *Two Essays on Entropy*, Berkeley: University of California Press.
1980: "A Basic System of Inductive Logic, Part II," in *Studies in Inductive Logic and Probability*, ed. R. C. Jeffrey, vol. 2, Berkeley: University of California Press, pp. 8–155.

1971 (with Jeffrey, R. C.) (eds.): *Studies in Inductive Logic and Probability*, vol. 1, Berkeley: University of California Press.

Works by other authors

Ayer, A. J. (1936) *Language, Truth and Logic*, London: Gollancz.
Church, A. (1956) *Introduction to Mathematical Logic*, vol. 1, Princeton, NJ: Princeton University Press.
Friedman, M. (1992) "Epistemology in the *Aufbau*," *Synthese* 93, pp. 191–237.
Gödel, K. (1931) "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme 1," *Monatshefte für Mathematik und Physik* 38, pp. 173–98.
Goodman, N. (1951) *The Structure of Experience*, Cambridge, MA: Harvard University Press.
Haack, S. (1977) "Carnap's *Aufbau*: Some Kantian Reflections," *Ratio* 19, pp. 170–5.
Hempel, C. G. (1950) "Problems and Changes in the Empiricist Criterion of Meaning," *Revue Internationale de Philosophie* 11, pp. 41–63.
Hilpinen, R. (1975) "Carnap's New System of Inductive Logic," in *Rudolf Carnap, Logical Empiricist: Materials and Perspectives*, ed. J. Hintikka, Dordrecht: Reidel, pp. 333–59.
Jeffrey, R. C. (ed.) (1980) *Studies in Inductive Logic and Probability*, vol. 2, Berkeley: University of California Press.
Kaplan, D. (1975) "Significance and Analyticity: A Comment on Some Recent Proposals of Carnap," in *Rudolf Carnap, Logical Empiricist: Materials and Perspectives*, ed. J. Hintikka, Dordrecht: Reidel, pp. 87–94.
Keynes, J. M. (1921) *A Treatise on Probability*, London: Macmillan.
Lewis, C. I. (1918) *A Survey of Symbolic Logic*, Berkeley: University of California Press.
Quine, W. V. (1951) "Two Dogmas of Empiricism," *Philosophical Review* 60, pp. 20–43.
Richardson, A. (1998) *Carnap's Construction of the World*, Cambridge: Cambridge University Press.

Roozeboom, W. (1960) "A Note on Carnap's Meaning Criterion," *Philosophical Studies* 11, pp. 33–8.
Sarkar, S. (1992) " 'The Boundless Ocean of Unlimited Possibilities': Logic in Carnap's *Logical Syntax of Language*," *Synthese* 93, pp. 191–237.
Sauer, W. (1985) "Carnap's '*Aufbau*' in Kantianischer Sicht," *Grazer Philosophische Studien* 23, pp. 19–35.
Shimony, A. (1992) "On Carnap: Reflections of a Metaphysical Student," *Synthese* 93, pp. 261–74.
Stegmüller, W. (1972) "Homage to Rudolf Carnap," in *PSA 1970*, ed. R. C. Buck and R. S. Cohen, Dordrecht: Reidel, pp. lii–lxvi.
Stein, H. (1992) "Was Carnap Entirely Wrong, After All?," *Synthese* 93, pp. 275–95.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

7

Karl Popper (1902–1994)

W. H. NEWTON-SMITH

Born in Vienna, Karl Popper studied at the University of Vienna from 1918 to 1922, after which he became apprenticed to a master cabinetmaker, Adalbert Posch. In his intellectual autobiography, Popper reported that he learned more about epistemology from Posch than from any other of his teachers. In 1925 he enrolled in the City of Vienna's new Pedagogic Institute to work on the psychology of thought and discovery. However, his interests turned to methodology and in 1928 he obtained his doctorate for a thesis on methodological problems in psychology.

While teaching mathematics and physics in a secondary school he wrote his *Logik der Forschung*, which was published in 1934, appearing in an English translation in 1959 as *The Logic of Scientific Discovery*. In 1937 he went to New Zealand as a lecturer in philosophy at Canterbury University College. While there he wrote his influential works *The Poverty of Historicism* and *The Open Society and its Enemies*. Appointed Reader and subsequently Professor in Logic and Scientific Method at the London School of Economics in 1946 he remained there for the balance of his academic career. Until his death in 1994 he continued to publish prolifically.

The distinctive feature of Popper's philosophy of science is his attitude to induction. Like Hume he held that no inductive inference is ever rationally justified. Finding that 1 million randomly selected samples of sodium burn with a yellow flame provides no reason at all, according to Hume and Popper, for thinking that all pieces of sodium will burn with a yellow flame. What we would normally count as evidence for such a hypothesis does not even give fallible grounds for thinking it is more probably true than false. Inductive arguments, arguments in which the premises do not entail the conclusion but purport to support it, simply have no rational force. Consequently Popper sought to rely entirely on deductive argumentation. While we can never have the least positive reason for thinking that a hypothesis is true or probably true, we can use a deductive argument to show that it is false. For given that we have observed one black swan we can deductively infer that it is false that all swans are white. This is the crux of Popper's philosophy of science. It is only the rejection of beliefs or hypotheses that can have the sanction of reason (but see below).

Hume never sought to persuade us to abandon induction. For him, it is part of our nature to proceed inductively. Custom and habit carry us forward where reason fails. To put the point anachronistically, for Hume we are "hard-wired" to induct. It is

simply a bemusing feature of the human condition that our inductive procedures do not have the sanction of reason. But for Popper, on the other hand, we do not or should not proceed inductively. And he claims that good scientists never do so. The Popperian scientist, equipped with a fertile imagination, simply makes a bold conjecture and the bolder the better. He then seeks to refute that conjecture by observation and experimentation. If a contrary instance is found the conjecture is falsified and hence rejected. In which case the scientist starts again with a new conjecture. If a conjecture is not falsified in a test, it has been "corroborated." Corroboration, as defined by Popper, does not provide any reason for thinking that the hypothesis has any likelihood of holding in the future. It is simply a report that it has not yet failed. Critics have wondered why, in this case, we should trust even a highly corroborated hypothesis. Clearly we do so when, for example, we trust our fate to airplanes designed on the basis of aeronautical theories. The answer for Popper is that we have no reason at all to do so. We proceed on blind faith! Critics also object that rejecting a hypothesis in the face of a contrary instance is itself a disguised form of induction. For in so doing we are assuming that the future will be like the past: what failed on Monday will also fail on Tuesday.

In utterly rejecting anything other than deductive justification, Popper committed himself to a very strong form of fallibilism, according to which not only can we not have certain knowledge in science or in everyday life, we can have no positive reasons, however weak, for holding that particular beliefs in science or in everyday life are even more likely to be true than false. Some readers of Popper may well have failed to see the extreme consequences of his fallibilism. For it applies also to the beliefs we form about what we observe. Consequently, it follows that we can have no rational grounds for claiming to have discovered that a hypothesis had been falsified. For we can have no more reason for rejecting a hypothesis than we have for our belief that we have observed a counterexample. That being so, Popper's fallibilism amounts to an extreme form of skepticism. We can have no reasons for thinking that any empirical proposition is true; nor can we have any reasons for thinking that it is false. Much of what Popper wrote has plausibility only if we set aside this extreme consequence of his position and this I will do in much of what follows.

Falsification provided Popper with his criterion for the demarcation of science from non-science or pseudo-science. He described this criterion as the very center of his philosophy of science. Impressed positively by the success of Einstein and negatively by what he took to be the failure of Freud and Marx, he looked for the hallmark of the scientific and thought he had found it in falsification. A theory is scientific just in case it makes predictions that could in principle be observed not to obtain. If they do not obtain the theory is refuted. According to Popper psychoanalysis ruled nothing out and hence could not be falsified and was not scientific. He held that Marxism was originally falsifiable. However, in the face of negative instances, Marxism was revised so as to become immune from refutation. In Popper's terminology any such unfalsifiable theories are metaphysical rather than scientific. Unfortunately, much of what we count as science turns out not to be falsifiable. A theory, such as quantum mechanics, which makes only probabilistic predictions, is not falsifiable. Consider the hypothesis that the probability that this coin will land heads on the next toss is p. The coin may land tails any number of times without falsifying that hypothesis! Popper sought to avoid this difficulty by adopting a methodological rule that would reject this hypothesis if after some number

N of trials the results diverged significantly from *p*. Critics have not been satisfied that there is any reasonable way of fixing an appropriate value for *N* without the use of inductive argumentation.

Popper did come to appreciate that there are metaphysical elements in good scientific theories and that metaphysical theories such as Darwinism had important beneficial influences on scientific development. Having recognized this Popper shifted somewhat and sought to evaluate metaphysical theories as well. This is to be done by considering whether the theory solves problems, whether its purported solutions can be examined critically and whether it solves the problems better than rival theories. This in turn generates a more general demarcation criterion for distinguishing between what he referred to as criticizable versus non-criticizable theories. It is no doubt a step forward to consider the merits of theories in general without special regard to whether they are scientific or not. But it means abandoning what once had pride of place within his philosophy of science: a hallmark of the scientific. Even Freud and Marx meet the condition of being criticizable theories, as Popper's own writings make manifest.

The method of science is to propose bold theories, the bolder the better. The scientist then seeks to refute them. Devoting oneself merely to finding out that theories are false does not seem a very edifying vocation. If there were only a finite number of theories in any branch of science, one could take comfort in the fact that with each rejection, the probability of the next theory selected being true would increase. But unhappily there are an infinite number of rival theories. A Popperian scientist expects that even his most cherished theory will eventually be falsified. If all he can ever find is that a theory is false, what positive gloss can he put on his scientific endeavors? For Popper the scientist hopes to have theories – false theories – that are ever better approximations to the truth. These are theories with, in his words, "increasing truthlikeness or verisimilitude." The move from Newton to Einstein was progressive, because while Newton said some true things about the world and some false things, Einstein said more true things and fewer false things. We can picture this as Newton getting a certain percentage of his claims right and Einstein scoring a higher percentage. Some future scientist can hope for a higher score yet. The idea that the aim of science is not truth *per se* but ever more approximately true theories has attracted adherents including many who reject Popper's account of scientific method. Unfortunately Popper's own technical definition of verisimilitude proved unsatisfactory. It turned out that on his definition all theories other than true ones have the same degree of truthlikeness. Popper's approach has inspired much further work on this notion but at present no satisfactory explication of truthlikeness has been forthcoming.

Popper is convinced that science is generating ever more truthlikeful theories. But as someone who avoids all inductive argumentation he has to regard this belief as irrational. Intuitively we might argue that the fact that Einstein passes more tests than Newton gives us reasonable grounds for thinking that Einstein's theory is more truthlikeful than Newton's. But this is an inductive argument. The conclusion is reasonable only if we assume that the area of the universe we have explored to date is a representative sample of the entire universe. Perhaps it is just a local peculiarity that Einstein fares better than Newton.

Popper himself is tempted by such arguments. At one point he claimed that we could argue for the greater truthlikefulness of Einstein over Newton on the grounds that it

would be "a highly improbable coincidence if a theory like Einstein's could correctly predict very precise measurements not predicted by its predecessors unless there is 'some truth' in it" (Schlipp 1974: 1192–3). This argument has many adherents particularly among those who advocate realism about scientific theories. But it is a form of induction known as *inference to the best explanation*. We are invited to infer that Einstein's theory has more truth in it than Newton's theory on the grounds that that assumption provides the best explanation of the greater predictive success of Einstein's theory. Respectable as this argument may be, it is not open to one who rejects all but deductive argumentation. One might feel that in all fairness we should allow Popper just this one little inductive move. He himself concedes in the passage quoted above that "there may be a 'whiff' of inductivism here." But if an inductive move is legitimate here, why not elsewhere as well? Once we allow induction a role, Popper loses claim to our attention. For what made his philosophy of science unique and thereby interesting was its explicit and total rejection of induction. But without induction, his belief in scientific progress is irrational.

Philosophers of science are divided on many issues. But they are almost unanimous in rejecting a Popperian account of science. Whether or not we have a satisfactory answer to Hume's skepticism about induction, it takes courage to deny that scientists proceed inductively. Scientists make an inductive move when they conclude that there is some probable truth in the theory of the electron on the grounds that that theory explains why televisions work. Even the scientist who concludes more modestly that there are at least good reasons to think that the theory of the electron will give successful observational predictions in the future is assuming the legitimacy of induction. Popper's falsificationist theory of science is itself falsified by scientific practice. His grand experiment in offering a non-inductivist theory of science serves only to heighten our appreciation of the deep-seated commitment to induction.

Popper argued for a number of philosophical positions quite independent of his falsificationism. For instance, he was a passionate defender of the freedom of the will. He argued against the determinist's thesis that the future is fixed by the past states of the universe together with the laws of nature. His strategy was to seek to show that not all future events involving human agents can be scientifically predicted. However, critics have been unable to see how an inability even in principle to predict some future human actions shows that those actions may not be determined nonetheless. More is needed to establish freedom than an inability to predict. In addition he has urged a controversial three-part ontology. He posits a world of physical objects ("world 1"), a world of subjective experiences ("world 2"), and a world of the "objective contents of thought" ("world 3"). "World 3" is reminiscent of the world of abstract Platonic objects that some philosophers have felt driven to postulate. But what is perhaps quite unique to Popper is the thought that this world, initially created by us, takes on an autonomy whereby it acts in a quasi-causal way on the objects of "world 2" and even of "world 1." Few philosophers have been willing to follow him in this lavish postulation of an unfalsifiable theory of causally active abstract objects.

History is likely to remember Popper more as a cultural figure than a philosopher in the narrow Anglo-Saxon sense of the term. For through his *The Poverty of Historicism* (1944) and his *The Open Society and its Enemies* (1945), he may well have had more influence in the social and cultural spheres than any other twentieth-century

philosopher. His initial aim was to provide an intellectually decisive critique of Marx and Marxism. The resulting polemic became a hugely influential attack on totalitarianism in general and a glimpse of an inspiring if only vaguely sketched Utopia: the Open Society.

Popper saw both fascism and Communism as resting on a pernicious *historicism*, a vision of history moving inevitably to some fixed final destination. Popper's historicist thinks he can detect by intuitive observation historical trends that he mistakenly takes to be iron laws and not mere trends that can be reversed. Popper characterizes Marxism, his paradigm of historicism, as bad historical prophecy combined with the injunction "*Help to bring about the inevitable!*" (1976: 35). Even granting Popper's assumption about the negative role that a belief in historicism has played and even granting the cogency of his arguments against it, his position does not really address the general problem of totalitarianism. For unfortunately there are totalitarian regimes that have come about as a result of forces other than belief in historicism.

Popper, in contrast to the historicists' policy of waiting for the inevitable, grand-scale, social change, advocates "piecemeal social engineering." We should make small experimental adjustments in our social institutions; this he illustrates with such examples as the introduction of a new sales tax. We then observe the results of the test; find out our errors and learn from our mistakes. For Popper this is explicitly an extension to the sphere of politics of the scientific method, although ironically the boldness that characterized the ideal scientist's conjectures is to be replaced by cautious small-scale conjectures. The bad moves are exposed; new ones are tried in their place. Our social engineer is urged to undertake a "systematic fight against definite wrongs, against concrete forms of injustice or exploitation, and avoidable suffering such as poverty or unemployment" (1957: 91).

In the social and political sphere, the notion of scientific rationality or reasonableness is liberalized to give a wider notion of rationality or reasonableness as openness to criticism: an attitude of readiness to listen to critical arguments and to learn from experience. A commitment to rationality in this sense is a necessary core of Popper's lightly sketched vision of an open society. An open society is a democratic one that promotes criticism and diversity without repression or irreconcilable social divisions, avoids violence, and encourages toleration. Critical public discussion with the participation of all is the means whereby those in an open society seek to arrive at a consensus on social and political issues. While the details are slight, Popper makes clear his touchingly naive view that the "free world" is a reasonable approximation to his ideal of an open society, having "very nearly, if not completely, succeeded in abolishing the greatest evils which have hitherto beset the social life of man" (1963: 370).

The core of the idea of an open society is supposed to lie in the recognition of our fallibility. Popper's approach is reminiscent of Mill, who argued in *On Liberty* that we have an interest in promoting diversity of opinion. Beliefs may be incorrect. If we are wrong, our best hope in correcting ourselves lies in critical discussion with those who disagree. And if we are in fact correct, the critical discussion will bring out more clearly the contents of and grounds for our beliefs. But it is hard to see how any argument of this character can give us Popper's full conception of an open society. For instance, an interest in critical discussion is logically compatible with a majority maintaining a class of slaves whose role in life was to provide us with critical comment! Attractive as the

nexus of values embodied in the open society are, they do not seem derivable just from the refutation of historicism together with the recognition of our fallibility. While society would no doubt benefit from more critical dialogue, it is arguable that the stability of social institutions requires that not everything is open to debate at all times.

There are serious tensions between the method of piecemeal social engineering and the advocacy of an open society. On the one hand, this method may only be viable for those who have already established an open society. The transitions in Albania and Romania, for example, from a closed to an open society were not achieved by piecemeal social engineering and probably could not have been achieved by anything other than revolution. On the other hand, it is not clear that all problems facing open societies can be solved through piecemeal social engineering. Globalization presents problems the resolution of which will require international regulatory mechanisms and these will not come about through piecemeal social engineering.

For all the deficiencies in argumentation and conception, Popper's rhetoric has been much used and used to largely beneficial effect. *The Poverty of Historicism* and *The Open Society* were texts widely read in samizdat under Communism and certainly had an inspirational role for dissidents seeking to open their societies. At the same time these texts were widely used by western European social democratic parties of the left to distance themselves from the Communist Parties of Europe. Progress to social equality was to be achieved by piecemeal social engineering, not revolutionary change. And in China in the period just before Tiananmen Square the liberal wing of the Communist Party used Popper explicitly in their analysis of the mistakes of the Cultural Revolution and in advocating open, critical discussion of social issues.

Popper is no doubt right in encouraging more critical discussion of social and political issues. Unfortunately his own philosophical system, with which he seeks to underpin this encouragement, is limited. His utter rejection of induction, his fallibilism, severely restricts the scope of rational criticism. Given that only deductive argumentation has rational force, the only intellectually justified criticism that can be made of any position in science or in society is that the position is logically incoherent. But such a limited form of criticism is unlikely greatly to assist in the solution of the social ills that concerned Popper.

Bibliography

Works by Popper

1945: *The Open Society and its Enemies*, London: Routledge and Kegan Paul.
1957: *The Poverty of Historicism*, London: Routledge and Kegan Paul.
1959: *The Logic of Scientific Discovery*, London: Hutchinson.
1963: *Conjectures and Refutations*, London: Routledge and Kegan Paul.
1972: *Objective Knowledge*, Oxford: Oxford University Press.
1976: *Unended Quest: An Intellectual Autobiography*, London: Collins.

Works by other authors

Newton-Smith, W. H. (1981) *The Rationality of Science*, London: Routledge and Kegan Paul, ch. III.
O'Hear, A. (1980) *Karl Popper*, London: Routledge and Kegan Paul.

O'Hear, A. (ed.) (1995) *Karl Popper: Philosophy and Problems*, Cambridge: Cambridge University Press.
Raphael, F. (1998) *Popper*, London: Phoenix.
Schilpp, P. A. (ed.) (1974) *The Philosophy of Karl Popper*, La Salle, IL: Open Court.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

8

Gilbert Ryle (1900–1976)

AVRUM STROLL

Gilbert Ryle and his junior colleague, J. L. Austin, were the leading figures of post-World War II Oxford philosophy. Though their aims and methods were different (see below), both are correctly characterized as "ordinary language philosophers." Unlike Austin, who published only seven papers in his lifetime, Ryle was a prolific writer. Much of what we know about his personal life derives from self-references in his numerous biographical sketches and reviews, and especially from his autobiography. In these various essays he describes his interactions with, and assessments of, the foremost philosophers of the time, among them Wittgenstein, Moore, Collingwood, Carnap, Prichard, H. H. Price, and Austin. His autobiography is to be found in *Ryle*, edited by O. P. Wood and G. Pitcher (1970). Although it is only fifteen pages long, it is wittily self-deprecating, devastating in its depiction of the state of philosophy in Oxford in the 1920s and 1930s, packed with information, and instructive with respect to his philosophical development. About Oxford philosophy he says:

> During my time as an undergraduate and during my first years as a teacher, the philosophical kettle in Oxford was barely lukewarm. I think that it would have been stone cold but for Prichard, who did bring into his chosen and rather narrow arenas vehemence, tenacity, unceremoniousness, and a perverse consistency that made our hackles rise as nothing else at that time did. The Bradleians were not yet extinct, but they did not come out into the open. I cannot recollect hearing one referring mention of the Absolute. The Cook Wilsonians were hankering to gainsay the Bradleians and the Croceans, but were given few openings. Pragmatism was still represented by F. C. S. Schiller, but as his tasteless jocosities beat vainly against the snubbing primnesses of his colleagues, even this puny spark was effectually quenched . . . Soon Oxford's hermetically conserved atmosphere began to smell stuffy even to ourselves.

About himself he states that in his mid-twenties he decided that philosophy essentially involves argumentation, and therefore that "the theory and technology" of reasoning needed to be studied by any would-be philosopher. Since nothing of that sort was available in Oxford he "went all Cambridge," and seriously began to study Russell; but, as he frankly admits, with marginal qualifications:

> Having no mathematical ability, equipment or interest, I did not make myself even competent in the algebra of logic; nor did the problem of the foundations of mathematics

> become a question that burned in my belly. My interest was in the theory of Meanings – horrid substantive! – and quite soon, I am glad to say in the theory of its senior partner, Nonsense. I laboured upon the doublets – Sense and Reference, Intension and Extension, Concept and Object, Propositions and Constituents, Objectives and Objects, Facts and Things, Formal Concepts and Real Concepts, Proper Names and Descriptions, and Subjects and Predicates. It was in Russell's *Principles of Mathematics* and not in his *Principia Mathematica*, in his Meinong articles and his "On Denoting," that I found the pack-ice of logical theory cracking. It was up these cracks that Wittgenstein steered his *Tractatus*.

His interests in the theories of meaning and reference were to dominate the remainder of his career, and differentiate his version of ordinary language philosophy from Austin's. Austin's main concerns were in the utterances that constitute promises, warnings, recommendations, admonishments, counsels, and commands, – i.e. in so-called "speech acts" (see AUSTIN) – whereas Ryle saw his task as that of distinguishing locutions that make sense from those that do not. In a succinct passage Ryle explains the difference between his task and Austin's.

> An examiner might pose two questions:
>
> (1) Why cannot a traveller reach London gradually?
> (2) Why is "I warn you ..." the beginning of a warning, but "I insult you" not the beginning of an insult?
>
> On six days out of seven Question 1 would be Ryle's favourite; Question 2, Austin's. Each of us would think – wrongly – that there is not much real meat in the unfavoured question. But their meats are of such entirely disparate kinds that the epithet "linguistic" would apply in totally different ways (1) to the answer-sketch, "Adverbs like 'gradually' won't go with verbs like 'reach' for the following reason ..."; (2) to the answer-sketch "To insult is to say to someone else pejorative things with such and such an intention, while to warn is to say ..." Anti-nonsense rules govern impartially sayings of all types. "Reach gradually" will not do in questions, commands, counsels, requests, warnings, complaints, promises, insults, or apologies, any more than it will do in statements. Epimenides can tease us in any grammatical mood. To an enquiry into categorial requirements, references to differences of saying-type are irrelevant; to an enquiry into differences between saying-types, references to category-requirements are irrelevant. Infelicities and absurdities are not even congeners.

As Ryle points out these different approaches were not in competition, but rather represented two parallel paths that "informal philosophy" could legitimately take in dealing with philosophical problems. Among those who emphasized the sense/nonsense distinction were Wittgenstein, Moore, J. T. Wisdom, O. K. Bouwsma, and Norman Malcolm. Austin's focus on speech acts was later to influence the work of Paul Grice, Zeno Vendler, John Searle, and A. P. Martinich. And, of course, there are many philosophers, including Ryle and Austin and some of those just mentioned, in which both approaches play concurrent roles.

In the twenty years between 1927 and 1947, Ryle had published more than thirty articles, reviews, and critical notices, but no books. His first venture into this larger format was *The Concept of Mind* (1949). Apart from collections of his essays he was to publish only two other books in his lifetime, *Dilemmas* in 1954, and *Plato's Progress* in 1966. In the former book, Ryle discusses six tensions (dilemmas) that are not counter-

vailing formal *theories* but rather opposing "platitudes." Each is an analogue of a classical philosophical perplexity, such as the free will problem. Thus, "In card games and at the roulette-table it is easy to subside into the frame of mind of fancying that our fortunes are in some way prearranged, well though we know that it is silly to fancy this." Ryle shows by a subtle, piecemeal analysis of the linguistic idioms in which the opposing platitudes are framed how the apparent dilemma is factitious and can be dissolved. *Plato's Progress* is an entirely different kind of book. It is a historical analysis in which Ryle tries to give a different portrayal of Plato's career. It is a provocative treatise that questions the common view that Aristotle was Plato's pupil, and that gives new datings to the Platonic dialogues.

Though these monographs are exciting pieces, and well worth serious study, they do not match the power and depth of the *Concept of Mind.* It has two aspects: a negative, deflationary one and a positive, constructive one. The two approaches are tied together by an attack on a certain picture of the human mind and its relationship to the human body. Ryle gives different names to this picture: He calls it the "Official Doctrine," the "Cartesian Model," "Descartes' Myth," the "Ghost in the Machine," and the "Para-Mechanical Hypothesis." The negative attack is to show that this picture is incoherent; the positive contribution is to give an accurate account (not a picture) of the relationship between mind and body. The positive account is detailed. It deals with the entire range of the mental: the will, knowing, emotions, dispositions and occurrences, self-knowledge, sensation, observation, imagination, and the intellect. The book is thus a treasure-house of detailed descriptions of all the major features of mentation.

What is the Official Doctrine he is out to destroy? This doctrine, he contends, is given its canonical formulation by Descartes, but its antecedents are much older. It is widely accepted by philosophers, psychologists, religious teachers, and many ordinary persons. It holds that every human being is both a mind and a body that are ordinarily harnessed together, but that after the death of the body the mind may continue to exist and function. Human bodies are in space and are subject to the mechanical laws of physics, chemistry, and biology. The body is a public object and can be inspected by external observers. But minds are immaterial, and are not in space, nor are their operations subject to mechanical laws. The mind is an entity, to be sure, but an immaterial and invisible one that inhabits a mechanical body. This is why Ryle calls it the "ghost in the machine." It is *res cogitans* in Descartes' parlance. It is the thing that thinks, deliberates, decides, wills, and opines. Each mind is private, i.e. only each person can take direct cognizance of the states and processes of his or her own mind.

A person thus lives through two collateral histories; one consisting of what happens to his body, the other to what happens within his mind. The first is public, the second private. The Cartesian picture thus depends on the internal/external distinction. This leads to the problem of how the mind influences bodily action. Since the mind is construed as nonphysical and nonspatial how does one's act of will, say, lead to a movement of one's legs, i.e. to the sort of thing called walking, for instance? Moreover, how are we to account for the knowledge we presume we have of the minds of others? If the Cartesian model is correct, observers cannot know what is taking place in the mind of another, since they are in principle cut off from any sort of direct cognitive awareness of that person's mental states or processes. The only direct knowledge any human has is of his or her own mental functions.

As plausible as this view may seem, it is absurd according to Ryle. It is one big mistake and a mistake of a special kind that he calls a "category mistake." To illustrate what he means by a "category mistake," Ryle offers three examples. Here is an abbreviated version of the first of these:

> A foreigner visiting Oxford or Cambridge for the first time is shown a number of colleges, libraries, playing fields, museums, scientific departments and administrative offices. He then asks "But where is the University? I have seen where the members of the Colleges live, where the Registrar works, where the scientists experiment and the rest. But I have not yet seen the University in which reside and work the members of your University." . . . His mistake lay in his innocent assumption that it was correct to speak of Christ Church, the Bodleian Library, the Ashmolean Museum *and* the University, to speak, that is, as if "the University" stood for an extra member of the class of which these other units are members. He was mistakenly allocating the University to the same category as that to which the other institutions belong. (1949: 16)

Ryle's point is that this sort of mistake is made by people who do not know how to employ the concept of a university. That is, their puzzle arises from an inability correctly to use certain items in the English vocabulary. According to Ryle, the Official Doctrine arises from a category mistake analogous to the preceding. It assumes that minds belong to the same category as bodies in the sense that both are rigidly governed by deterministic laws. The human body works according to mechanical principles: the heart is a pump, the veins are pipes, and the flow of blood is determined by the pressures that are described in fluid mechanics. The system is thus an assemblage of interacting parts that consist of fluids, solids, and electrical forces, all of which operate according to the laws of mechanics. All these forces usually work to some desired end, such as moving blood from one part of the body to another.

Minds also work in analogous ways. When I am hungry, a mental state, a desire, acts on my body and initiates those movements of hands and fingers that allow me to pick up and transfer food to my mouth. Accordingly minds must be governed by deterministic laws. But minds are nonmaterial. They are not composed of solids, fluids, and electrical forces. So their laws, though deterministic, are non-mechanical. These Ryle calls "para-mechanical." The Official Doctrine invokes them as the analogues of the mechanical laws that govern the behavior of physical entities. But the concept of a para-mechanical law is absurd. There are no such things as immaterial levers, valves, and pumps. Valves, levers, and pumps are solid entities that operate to effect physical movements. To invoke the immaterial analogues of such entities to explain mental activity is thus to make a category mistake, i.e. to apply the concepts of mechanical forces and laws to a domain where they have no grip. The mistake arises because philosophers do not know how to employ the ordinary epithets we use for describing mental activity. Philosophers are thus like the person who does not know how to employ the concept of a university. It is this para-mechanical model that Ryle attacks in his book. Its existence indicates that these theorists do not know how to wield the set of concepts that characterize our mental functions.

The alternative he offers to the Official Doctrine is a detailed description of how mental concepts are used in everyday life. As he says: "The philosophical arguments

which constitute this book are intended not to increase what we know about minds but to rectify the logical geography of the knowledge which we already possess." Ryle is thus reminding us of what we have always known, and also reminding us how philosophical conceits can blind us to the familiar. His description of the "logical geography" of mental concepts is thus a reminder of how we employ these concepts when we are not doing philosophy. Since any such employment is enormously complex, its "logical geography" will be lengthy, detailed, and specific. Here, by way of illustration, is a segment of a much longer specimen of logical geography:

> It is true that the cobbler cannot witness the tweaks that I feel when the shoe pinches. But it is false that I witness them. The reason why my tweaks cannot be witnessed by him is not that some Iron Curtain prevents them from being witnessed by anyone save myself, but that they are not the sorts of things of which it makes sense to say that they are witnessed or unwitnessed at all, even by me. I feel or have the tweaks, but I do not discover or peer at them; they are not things that I find out about by watching them, listening to them, or savouring them. In the sense in which a person may be said to have had a robin under observation, it would be nonsense to say that he has had a twinge under observation. There may be one or several witnesses of a road-accident; there cannot be several witnesses, or even one witness, of a qualm. (1949: 205)

This passage is a good example of Ryle's way of exorcizing the ghost in the machine. The Official Doctrine presupposes that one has privileged access to a private realm consisting of one's own sensations, thoughts, and mental states; and that such an access consists in the observation of one's sensations and states. But to say that one is observing something implies that one is using one's eyes, or certain kinds of observational aids such telescopes, stethoscopes, and torches. One's eyes, and these instruments, can be used for the observation of planets, heart-beats, and moths. But we do not know what it would be like to apply them to felt sensations or to assert seriously that we "observe our pains." Since the Official Doctrine presupposes there is such a para-mechanical analogue as observing, it can be shown to be a species of nonsense by comparing its requirements with our actual use of such mental concepts as "tweaks" and "qualms." What the comparison reveals is a category mistake. The concept of observation applies to the physical domain in a way it *logically cannot* apply to the mental. Just as one logically cannot reach London gradually, so one cannot "observe" one's aches and pains. Ryle's line of reasoning throughout the work is thus to show that theorists have incorrectly wielded the ordinary concepts that describe human mental life.

The Concept of Mind created a sensation when it appeared in 1949. For at least a decade after its publication it was the single most discussed book in Anglo-American philosophy. Nearly every periodical carried long articles about it. It was translated into a host of foreign languages, was taught in virtually every major western university, and within a short time seemingly had achieved the status of a philosophical classic. Yet a decade later it had fallen into obscurity, and subsequently it has hardly been referred to at all. What happened to occasion such a collapse? It is especially puzzling given that the book was of superb philosophical quality, was elegantly written, introduced many original and powerful distinctions, and was the first study to show in detail how the

philosophy of language and the philosophy of mind are tied together. In this last respect, it was a bellwether for work that was to be developed thirty years later.

There are several possibilities to explain what happened. One factor is that four years later Wittgenstein's posthumous *Philosophical Investigations* appeared (see WITTGENSTEIN). It covered much the same territory as Ryle's study and in greater depth. As brilliant as Ryle's book was it paled in comparison to the power and insight of Wittgenstein's. So philosophers turned from Ryle to Wittgenstein. It was the latter and not the former who was now read: Ryle had simply gone out of fashion.

There is a second factor. Ryle claimed that in this work he was "charting the logical geography" of the many concepts used in speaking about the human mind. And though this was clearly an apt description it was also patent that his work had a strong verificationist thrust. Ryle frequently and in crucial passages speaks about the testability of propositions about mental concepts. For example, he states: "For, roughly, the mind is not the topic of sets of untestable categorical propositions, but the topic of sets of testable hypothetical and semi-hypothetical propositions" (p. 46). Some critics have thus emphasized that Ryle's aim is to correct what other philosophers have said about the methods of *verifying* statements involving mental concepts, rather than trying to explicate these concepts themselves. The positivists, of course, identified the meaning of a statement with the method of its verification, and in many places in the *Concept of Mind* Ryle seems to presuppose that in describing how certain propositions involving mental concepts are to be tested he is explicating the meaning of those concepts. The book was thus eventually assessed as a sophisticated form of logical positivism, a view which had lost its influence by the 1950s. Ryle's work was swept away with the rest of this movement.

Its behaviorism was a third factor. Ryle states that to give reasons for accepting or rejecting statements containing mental concepts will always involve hypothetical statements about overt behavior. In responding to the question, "What knowledge can one person get of the workings of another mind?" Ryle answers that it is "how we establish, and how we apply, certain sorts of law-like propositions about the overt and the silent behavior of persons. I come to appreciate the skill and tactics of a chess player by watching him and others playing chess" (p. 169). Although Ryle always denied that he was reducing mind to behavior, and asserted instead that charting the "logical geography" of mental concepts was a philosophically neutral endeavor, his detailed analyses seemed to many philosophers to leave out one fundamental characteristic of the mind, the inward, felt quality of mental experience. For these philosophers such mental activities as deliberating or conjecturing, or such states as being in pain, were distinct from behavior. One could, for example, be in pain without evincing it in any mode of behavior. And even if one were to evince it, the pain itself was not to be identified with the behavior in question. A pain is not a grimace. So even if Ryle were correct in arguing that mental activity was exercised in various intersubjective situations it did not follow that the behavior so exhibited was identical with the mental events in question. Unlike Ryle, who minimized internal experience, Wittgenstein emphasized and acknowledged the existence of such phenomena. His point was that one should not identify them with such features as meaning, expecting, thinking, and so forth. And this position was seen to be more compelling than Ryle's. In the end this may have been the decisive factor in the eclipse of Ryle's reputation.

Bibliography

Works by Ryle

1931–2: "Systematically Misleading Expressions," *Proceedings of the Aristotelian Society* 32, pp. 139–70.
1945: "Philosophical Arguments," Inaugural Lecture as Waynflete Professor of Metaphysical Philosophy, Oxford. (Reprinted in Ryle 1971, vol. 2, pp. 194–211.)
1945–6: "Knowing How and Knowing That," *Proceedings of the Aristotelian Society* 46, pp. 1–16.
1949: *The Concept of Mind*, London: Hutchinson.
1950–1: "Heterologicality," *Analysis* 11, pp. 61–9.
1954: *Dilemmas, The Tarner Lectures*, Cambridge: Cambridge University Press.
1961: "Use, Usage, and Meaning," *Proceedings of the Aristotelian Society*, supp. vol. 35, pp. 223–30.
1966: *Plato's Progress*, Cambridge: Cambridge University Press.
1971: *Collected Papers*, vols. 1 and 2, London: Hutchinson.

Works by other authors

Lyons, W. E. (1980) *Gilbert Ryle: An Introduction to his Philosophy*, Brighton: Harvester.
Wood, O. P. and Pitcher, G. (eds.) (1970) *Ryle*, Garden City, NY: Anchor Books.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

9

Alfred Tarski (1901–1983), Alonzo Church (1903–1995), and Kurt Gödel (1906–1978)

C. ANTHONY ANDERSON

Alfred Tarski

Tarski, born in Poland, received his doctorate at the University of Warsaw under Stanislaw Lesniewski. In 1942, he was given a position with the Department of Mathematics at the University of California at Berkeley, where he taught until 1968.

Undoubtedly Tarski's most important philosophical contribution is his famous "semantical" definition of truth. Traditional attempts to define truth did not use this terminology and it is not easy to give a precise characterization of the idea. The underlying conception is that semantics concerns *meaning* as a relation between a linguistic expression and what it expresses, represents, or stands for. Thus "denotes," "designates," "names," and "refers to" are semantical terms, as is "expresses." The term "satisfies" is less familiar but also plausibly belongs in this category. For example, the number 2 is said to *satisfy* the equation "$x^2 = 4$," and by analogy we might say that Aristotle satisfies (or satisfied) the formula "x is a student of Plato."

It is not quite obvious that there is a meaning of "true" which makes it a semantical term. If we think of truth as a property of sentences, as distinguished from the more traditional conception of it as a property of beliefs or propositions, it turns out to be closely related to satisfaction. In fact, Tarski found that he could define truth in this sense in terms of satisfaction.

The goal which Tarski set himself (Tarski 1944, Woodger 1956) was to find a "materially adequate" and formally correct definition of the concept of truth as it applies to sentences. To be materially adequate a definition must "catch hold of the actual meaning of an old notion," rather than merely "specify[ing] the meaning of a familiar word used to denote a novel notion" (Woodger 1956: 341). Again, in discussing the material adequacy of some of his other definitions, Tarski writes, "Now the question arises of whether *the definitions just constructed* (the formal rigour of which raises no objection) *are also adequate materially*; in other words *do they in fact grasp the current meaning of the notion as it is known intuitively?*" (Woodger 1956: 128–9).

To determine whether or not a proposed definition of a certain concept is materially adequate, Tarksi thinks that we must first formulate a *criterion* of material adequacy for such a definition: a precise condition which the definition must meet and which will guarantee that the defined notion is faithful to the original intuitive conception. Of

course, whether a proposed condition really guarantees sufficient conformity to the old notion is subject to critical review.

The requirement of *formal correctness* means that the proposed definition must be non-circular and that it must meet other logical constraints on acceptable definitions. One of the traditional requirements is that a definition must not define something in terms of things which are less clear than it. Tarksi even maintains that it must be specified which previously adopted terms are to be used in giving the definition and requires that the formal structure of the language in which the definition is to be given be precisely described.

These are rigorous constraints. The motivating idea seems to be that only under such conditions can we hope to *prove* the material adequacy and formal correctness of a definition of truth.

Tarski proposes as a criterion of material adequacy for a definition of truth that the definition shall have as logical consequences all instances of Schema (T):

(T) X is true if and only if p,

where "X" is replaced by a name of an arbitrary sentence of the language in question and "p" is replaced by that very sentence (or by a sentence with exactly the same meaning). The name in question must be a quotation-mark name or at least a name which necessarily designates the sentence. An appropriate instance of Schema (T) is thus such a thing as:

(S) "Snow is white" is true if and only if snow is white.

On the left-hand side of this "if and only if" there occurs a name of a certain sentence – which name is constructed by enclosing the sentence in question in quotation marks. Then using that name to *mention* the sentence, the property of being true is predicated of the sentence. On the right-hand side of the equivalence, the very sentence, which is named on the left and said there to be true, is *used*. The thing may appear to be a triviality and perhaps that is all to the good. The condition, after all, is supposed to constrain an adequate definition in such a way that satisfying this condition guarantees that the definition catches hold of the actual meaning of the term "true."

Note carefully that Schema (T) is *not* Tarski's definition of truth. That a definition should imply all instances of Schema (T) is the criterion of adequacy for the definition. But Tarski does seem to think that all the instances of (T) together completely capture the meaning of "true." If we could form an infinite conjunction, connecting all the instances with "and," we would have a complete specification of the semantical conception of truth. This is not an acceptable procedure according to the usual rules of definition, but a correct definition would be obtained if we could somehow achieve the same effect.

Now the conditions which have already been given for an acceptable definition of truth require that the language involved be specified quite precisely. Natural languages do not have, or at least we do not know, rules which determine exactly what its expressions are; for example, the sentences of English are not precisely specified. If we ignore this and set as our task to give a definition of truth for a natural language, say English, we encounter a paradox. No predicate of a sufficiently expressive language such as English can have the property that it validates every instance of Schema (T). And this

is so whether the predicate is defined or not. The proof of this appeals to the infamous liar antinomy (or paradox). In a very simple version the antinomy (something "contrary to law") goes like this. Consider

(A) A is not true.

That is, consider the sentence "A is not true," which sentence we have decided to name "A." Now Schema (T) implies:

(1) "A is not true" is true if and only if A is not true.

But observing that the sentence A is the very sentence "A is not true", we may assert:

(2) A = "A is not true."

If two things are identical, then they share all the same properties. So, substituting the left-hand side of (2) for the right-hand side in (1), we get:

(3) A is true if and only if A is not true.

In the propositional calculus, this has the form:

(4) $P \equiv \sim P$,

"P if and only if not-P" and this is equivalent to the explicit contradiction:

(5) $P \,\&\, \sim P$,

"P and not-P."

Something must give. If we are unwilling to give up the usual laws of logic, since (2) is undeniable, it appears that we must alter or modify Schema (T), our criterion allegedly determined by the very meaning of "true."

Tarski concludes, somewhat hastily, that ordinary language is inconsistent. The concept of truth must conform to Schema (T), but if we have such sentences as A, we arrive at a contradiction. The problem, says Tarski, is that natural languages are *semantically closed*, that is, they contain within themselves the terms and machinery for doing their own semantics. For example "is true in English" is itself a predicate of English. We must, he says, give our definition of truth in a *metalanguage* for the language whose sentences are in question. A metalanguage is a language which we may use to talk about another language. For example, in a book written in English which explains the grammar and meaning of the German language, the metalanguage is English. The language being studied is called the *object language*: in the case of this example, German. Further, claims Tarski, we must confine our attention to formalized languages which, unlike natural language, need not be semantically closed and which are otherwise precisely specified.

With these provisos, Tarski proceeds to show that definitions of truth can be given for object languages which do not contain semantical terms. His method of definition has the striking quality that the definition, given in a metalanguage, *does not itself use any semantical terms*. Because of the liar antinomy and other conundrums involving semantical notions, Tarski considered it important to give the definition in such a way that *no* semantical terms are presupposed as primitive or understood without definition.

To see how the definition would be given for a very simple formalized language, let us suppose that we have just two predicates: "R," meaning *is red*, and "S," meaning *is square*. In addition, suppose that the language contains a variable x; a sign for negation, say "–"; for conjunction, "&"; and a notation for universal quantification, "∀" meaning *For every*. Thus, for example, we can write "$\forall x - S(x)$" for "For every object x, x is not square" or, more naturally, "Nothing is square."

We assume that our metalanguage contains the means of expressing at least the very same notions as the object language. Here we are using a bit of English as metalanguage so that we have the words "is red" to mean the same as the predicate "R" in the simple formalized language. Now let some domain of objects be selected as the collection of things we will be talking about. One can then define *satisfaction* for the object language:

(1) An object satisfies "$R(x)$" if and only if it is red.
(2) An object satisfies "$S(x)$" if and only if it is square.
(3) An object satisfies a negation $\ulcorner -\phi \urcorner$ if and only if it does not satisfy ϕ.
(4) An object satisfies a conjunction of the form $\ulcorner \phi \,\&\, \psi \urcorner$ if and only if it satisfies ϕ and it satisfies ψ.
(5) An object satisfies a universal quantification $\ulcorner \forall \times \phi \urcorner$ if and only if every object (in the domain) satisfies ϕ.

Here ϕ and ψ are *formulae* of the formalized language. These are expressions which we have not really defined but which include such things as "$R(x)$" ("x is red"), "$-[S(x)$ & $R(x)]$" ("It is not the case that x is square and x is red"), as well as sentences such as "$-\forall x - S(x)$" ("It is not the case that for every x, x is non-square," i.e. "Something is square").

This doesn't look like a definition, but in fact it really does completely explain the meaning of "satisfies" as it applies to our simple language. Using these definitional rules on complicated expressions we can proceed step by step to simpler expressions until we get down to cases covered by (1) and (2). And it may look as if we have some kind of vicious circularity. For example, we have used "and" (in the metalanguage) to define satisfaction for expressions (of the object language) containing "&." But the appearance is deceptive. We have assumed that whatever we can say in the object language, we can say in the metalanguage, but not necessarily vice versa. This assumption does not introduce any logical or philosophical difficulty into the definition.

Finally we define truth:

A sentence ϕ is true if and only if every object (in the domain) satisfies it. Again, we haven't really defined the sentences of our object language, but they will be expressions in which no occurrences of the variable are "dangling." For example, "$-\forall x - [R(x)$ & $S(x)]$" ("Something is red and square") is a sentence, as opposed to a formula such as "$R(x)$" ("x is red"). Here the variable x is just a placeholder, indefinitely indicating something or other, but no definite thing.

It is not obvious that this definition actually conforms to the criterion of material adequacy. But it does. It can be proved that every instance of Schema (T), confined to sentences of our object language, is a consequence of this definition. The whole thing may seem trivial, but it is really quite amazing that in an appropriate metalanguage truth can be defined without appealing to any semantical notions. This means that it

has been defined in terms of things which are clearer: they are just the concepts of logic together with the concepts of the object language.

It remains to mention Tarski's work on the notion of *logical consequence* (Woodger 1956: 409–20). This, like the notion of truth, was used in an intuitive way by logicians and philosophers before Tarski, but it was the latter who made the notion precise.

Consider once again our simple formalized language. Do not select a particular domain and particular meanings for "S" and "R." Rather, contemplate any arbitrary *interpretation* of the language – any domain of objects whatsoever and any appropriate meanings for these symbols. The logical symbols "–," "&," and so on, are to retain their original meanings throughout.

For any such specification, we can explain *truth under that interpretation* along the lines used above for the particular interpretation we were considering. Suppose that in some interpretation a particular sentence, say "$\forall x[\mathrm{S}(x) \,\&\, \mathrm{R}(x)]$," comes out true. Then in that interpretation certain other sentences will come out true as well. For example, "$\forall x\mathrm{S}(x)$" and "$\forall x\mathrm{R}(x)$." In fact, this will always happen. If an interpretation makes our example sentence true, that interpretation will also make these two sentences true. In such a case, Tarski says that the latter two sentences are *logical consequences* of the first sentence. In general, a sentence ψ is a logical consequence of a sentence ϕ if and only if every interpretation which makes ϕ come out true also makes ψ come out true. And a sentence is defined to be a logical consequence of a collection, or set, of sentences if every interpretation which makes every sentence in the set come out true also makes the sentence in question come out true. Finally, Tarski defines a sentence to be *logically valid* if it comes out true under every interpretation.

The importance of such a definition is that we can now strictly define what it is for something to be a valid argument in our language. And, of course, the study of valid arguments is at the very heart of the discipline of logic. Using these definitions we can then prove that certain systems of logical rules are "complete" in the sense of being adequate to their intended purpose of capturing all valid inferences expressible in the language. For example, certain formulations of first-order logic, the logic of such notions as *and, not, or, if, . . . then, not, some all*, and the like were proved complete by Kurt Gödel, to be discussed below.

These two things, his definition of the concept of truth for formalized languages and his explication of the concept of logical consequence are Tarski's distinctive philosophical contributions. They are substantial indeed.

Alonzo Church

In 1927, Church received his Ph.D. from Princeton, where he taught from 1929 to 1967. Thereafter, he taught at UCLA until 1990. He was a long-time editor of the *Journal of Symbolic Logic*, which he helped to found. Church's philosophical contributions largely concern questions about the foundations of logic and mathematics, especially their ontology, and topics in the philosophy of language and in the related area of intensional logic.

Church's thesis is a hypothesis concerning the identification of the mechanically computable or calculable functions discussed below in connection with Gödel's incompleteness theorem. Church proposed as a precise mathematical analysis of the idea of

such functions that they be identified with the *lambda-definable* functions. This latter notion is too technical to be explained in detail here. Alan Turing independently proposed the identification of the mechanically computable functions with functions computable in principle by a precisely definable sort of abstract "machine," now called a *Turing machine*. This identification turned out to be equivalent to Church's thesis. That is, the class of lambda-definable functions is exactly the same as the class of functions computable by a Turing machine. Other attempts to analyze the notion in question have always led to the same class of functions. The identification of the class of mechanically ("algorithmically") calculable functions with the class of lambda-definable or Turing machine-computable functions (the "Church–Turing thesis") is now almost universally accepted.

Church's theorem, to be carefully distinguished from Church's thesis, is a theorem of mathematical logic to the effect that there is no effective (= mechanical) procedure for deciding whether or not a formula of first-order logic is valid.

Church was a Platonist or, as he preferred, a realist about the entities apparently described and studied by mathematics and logic. Numbers and other mathematical entities are, he believed, objectively existing, mind-independent objects and mathematics itself consists of truths about these things. Logic seems to require, if formulated in full generality, *propositions*, *properties*, and "*individual concepts*." These kinds of things, usually called *intensional entities*, are supposed to be abstract, real, and objective entities suitable to be the meanings of expressions in various languages. Propositions, for example, are claimed to be the meanings of declarative sentences, the same for synonymous sentences, whether in a single language or in two or more different languages.

Church's general methodological viewpoint about the formal sciences was a kind of "hypothetico-deductive rationalism." According to this view, intuitions or feelings of self-evidence provide initial support for assumptions about abstract entities. The theories of these are to be *formalized*, stated using the precise language and terminology of symbolic logic, and the results are to be evaluated using the sorts of criteria common to scientific procedures in general. One way we evaluate theories is by deducing consequences and thereby determining whether they are adequate to account for the data. In the formal sciences Church took the data to include the accepted facts of mathematics and logic.

Many of Church's philosophical contributions appear in reviews in the *Journal of Symbolic Logic*. His relatively few papers devoted explicitly to philosophical topics usually concerned questions about meaning and related topics in the philosophy of language. There are also arguments against nominalism as it is sometimes espoused in connection with mathematics, logic, or semantics.

As a sample of the latter (Church 1950), consider a nominalist attempt to give an analysis of certain statements apparently about propositions. Suppose it is claimed that such a sentence as (1) "Seneca said that man is a rational animal" is to be analyzed as: (2) "There is a language *S*′ such that Seneca wrote as a sentence of *S*′ words whose translation from *S*′ into English is 'Man is a rational animal'." This may already seem excessively complicated, but simpler attempts to analyze statements about assertion so that they concern such relatively concrete things as sentences are subject to easy refutation. To bring out clearly that (2) will not do as an analysis of (1), Church uses the "translation test," a procedure whose invention is usually attributed to C. H. Langford.

If we translate (1) into German, we get (1′) "Seneca hat gesagt, das der Mensch ein vernünftiges Tier sei". In translating (2) into German, note carefully that the word "English" must be translated as "Englisch" (not as "deutsch") and the quotation which forms part of (2) is to be translated as "Man is a rational animal" (not as "Der Mensch ist ein vernünftiges Tier"). This latter translation, call it (2′), certainly would not convey anything like the information which would be conveyed to a German speaker (who spoke no English) by (1′). Thus, argues Church, (1′) is not an acceptable analysis of (1). The basic idea of the objection, which can be seen even without appealing to translation, is that (1) does not say anything about any particular language (and so neither does its translation (1′)), whereas (2) makes specific reference to English.

A philosophical argument which has a quite surprising conclusion is given by Church (Church 1956: 24–5) as a more precise version of reasoning offered by Gottlob Frege. The conclusion of the argument is that sentences *denote* truth-values, true sentences denoting Truth (or The True) and false sentences denoting Falsehood (or The False)! Put like this, the thesis seems quite incredible, even unintelligible. Why suppose that sentences "denote" anything at all? And what, we may ask, are these alleged "objects," Truth and Falsehood? These are good questions, but the essential point of Church's argument (and Frege's before him) could be stated like this: the truth or falsity of a sentence is the only thing that stands to the sentence as the denotation of a (complex) name stands to its parts. To see this take such a sentence as (a) "Sir Walter Scott is the author of *Waverley*." If we replace "the author of *Waverley*" by an expression which denotes the same, "the man who wrote twenty-nine Waverley Novels altogether," we get a new sentence: (b) "Sir Walter Scott is the man who wrote twenty-nine Waverley novels altogether." If we are supposing that the "denotation" of a sentence, whatever it is, is unchanged if a denoting part is replaced by another with the same denotation, then this new sentence must have the same denotation as the original. Further, it is plausible (Church claims) that the sentence, (c) "The number, such that Sir Walter Scott wrote that many Waverley Novels altogether is twenty-nine," is so close in meaning to (b) as to have the same "denotation" (again, without yet assuming that we know what this is). But now let us replace the denoting expression "The number, such that Sir Walter Scott wrote that many Waverley Novels altogether" in (c) by an expression with the same denotation, namely; "The number of counties in Utah" (which is in fact twenty-nine). We then get a sentence which is supposed to have the same denotation as (c), (d) "The number of counties in Utah is twenty-nine" (again assuming that a sentence does not change its denotation if a denoting part is replaced by another with the same denotation).

Now compare our original sentence (a) "Sir Walter Scott is the author of *Waverley*" with (d) "The number of counties in Utah is twenty-nine." By the reasoning just explained, these two sentences must have the same "denotation." But the only meaning-relevant feature which they seem to have in common is that both are true. A little reflection on such examples points to the conclusion that the only thing that can be expected to remain invariant under such substitutions is the truth or falsity of the original sentence. So if "denotation" has an analog for sentences, it will have to be the *truth-values*, truth and falsity, which may be seen as mathematical abstractions. (Compare the mathematical abstraction of numbers, as objects, from collections or from concepts of collections.) The Church–Frege argument here may not be conclusive,

but the analogy uncovered is striking and it may well be a useful theoretical assumption for semantics that sentences "denote" truth-values. (See Anderson 1998 for further discussion.)

Church's most important philosophical ideas are contained in his work on the foundations of intensional logic (Church 1951, 1973, 1974). Philosophers and logicians contrast intension and extension, but it is by no means easy to give a clear characterization of these notions. In the case of sentences, Church would maintain that the sense, or intension, of the sentence is the proposition which it expresses and the denotation, as already explained, is the truth-value of the sentence. Logic as standardly taught in philosophy and mathematics departments makes no significant distinction between sentences with the same truth-value; arguments which turn on finer distinctions of meaning are simply not treated. Similar distinctions hold between the *set* of things of which a predicate is true, the extension of the predicate, and the property conveyed by the predicate, its intension. Again, a distinction between the meaning, strictly so-called, of an expression such as "The present president of the US" (its intension) and what it stands for, the actual person, is needed. Here we might say, again with Church, that the meaning of the expression in the strict sense is the *concept* that it expresses, its intension, but what it denotes or stands for, the person or, more generally, the object, is its extension.

So, as already explained, Church calls the proposition expressed by a sentence its *sense* and the truth-value that it stands for its *denotation*. Predicates have properties as their senses and sets as their denotations, and individual expressions (e.g. descriptive names) have certain concepts as their senses and what they stand for as their denotations. The relationship that holds between the sense of an expression and what it denotes let us call the *concept relation*, and symbolize it by the capital Greek letter Δ (delta). Then propositions are concepts of truth-values, properties are concepts of sets, and individual concepts are concepts of the individual things that the concepts characterize. Generalizing our terminology (as Church does), call anything that is capable of being the sense of some expression a *concept*.

The intensional logic that Church envisioned would have two kinds of intensional axioms: logical principles about Δ and principles that would specify the essential characteristics of propositions and other complex concepts. In connection with the latter, Church took it to be especially important to have axioms which give, or correspond to, criteria of identity for complex concepts. A criterion of identity in the present case is a principle that determines the identity or difference of the complex concepts expressed by different sentences (or predicates or descriptive names) in terms of some known relation between the sentences (complex expressions) themselves. An example would be the principle that two sentences express the same proposition if and only if they are logically equivalent; in our example, that is, they have the same truth-value necessarily, or on logical grounds alone.

We have already explained that a function of numbers is a correlation of a certain kind. Thus, *square* or *squaring* is said to be a function from numbers to numbers. In general, any correlation between the things in two collections is called a function. Generally, a function is just any conceivable correlation between the things in one collection and the things in another (or, possibly, the same) collection; it is allowed that two or more things in the first collection be correlated with the same thing in the second.

A name of a function has both a sense and (in general) a denotation. The sense is therefore a concept of the function denoted by that name. For example the expression "The squaring function" denotes the function that takes each number into its square and it has as its sense what is conveyed by "x^2." The combination of a name of a function with the name of something to which the function is applied (an *argument* of the function) will also have a sense: a complex sense involving the sense of the function name and the sense of the name of the entity to which the function is being applied.

The importance of this idea appears in the observation that *any* complex expression may be construed as being built up from a function expression, together with expressions for one or more arguments to which the function is applied.

Now let us write "$\Delta(X,Y)$" to mean that X is a concept of Y. The axioms which Church took to govern the delta-relation are:

(C1) For every X, Y, and Z, if $\Delta(X,Y)$ and $\Delta(X,Z)$, then $Y = Z$.

(C2) For every F and F_1, if $\Delta(F_1,F)$, then for every X and X_1, if $\Delta(X_1,X)$, then $\Delta(F_1X_1,FX)$.

(C3) For every F and F_1, if for every X and X_1, $\Delta(X,X_1)$ implies that $\Delta(F_1X_1,FX)$, then $\Delta(F_1,F)$.

(C1) says that anything which is a concept of something is a concept of exactly one thing. In (C2) and (C3), F is any function and FX is the result of applying that function to an argument X; that is, FX is the entity which is correlated with X by the function. Where F_1 and X_1 are concepts, we have just written "F_1X_1" for the complex concept that results when the concept F_1 is combined with the concept X_1. In these terms, (C2) amounts to the claim that if an expression denoting a function is combined with an expression denoting an argument (in some possible language), then the sense of the complex expression is the result of combining the sense of the function name with the sense of the argument name.

The proposed axiom (C3) is more problematic. To understand and accept it, one really must go along with a hypothesis that Church proposes to simplify the logic of the system. Church assumes that *a concept of a function can be taken to be a function from concepts to concepts*. This is fine for axiom (C2), which is then just understood in such a way that combining a concept of a function with a concept of an argument is nothing more than applying a certain kind of function to a certain kind of argument. But axiom (C3) is much bolder. It amounts to the claim that *any* function from concepts to concepts satisfying a certain condition is a concept of a certain function. It says: if a function applied to a concept of an argument always yields a concept of the output of some function applied to the argument thus concepted, then the function (from concepts to concepts) is a concept of the function from objects to objects.

This axiom leads to various difficulties, which cannot be explained here (see Anderson 1998). It is fair to say that even the basic principles of intensional logic, as Church conceived it, are still not settled.

Intensional principles of the second sort – those supposed to individuate complex concepts – are also still problematic. Church proposed three heuristic principles to guide the formulation of such axioms: (A) that logically equivalent expressions express the same concept, (B) that expressions that have exactly the same

syntactical structure and whose corresponding parts have the same meanings express the same concept, and (C) expressions that can be obtained from one another by applying the logical operation of lambda conversion express the same concept. Lambda conversion is a logically valid transformation of expressions, which we do not attempt to explain here.

The idea behind (A) works well if, but only if, one is dealing with reasoning involving no finer distinctions of meaning than are involved in arguments turning on *modality*: necessity, possibility, impossibility, and similar conceptions. Suggestion (C) appeals to the technical notion of lambda-conversion which is very difficult to motivate from a philosophical point of view. Hands down, the notion urged in (B) is the most promising. Church (who agreed with the assessment just offered) tried to implement this approach several times in his published work, but technical and logical difficulties still block the way of a satisfactory theory.

It is fair to say that this project to which Church contributed fundamental and important work, to establish a comprehensive and adequate general intensional logic, has not yet been completed. But his successful philosophical contributions are impressive indeed.

Kurt Gödel

Kurt Gödel received his doctorate in 1930 at the University of Vienna. He emigrated to the United States in 1940 and soon afterwards became a member of the Institute for Advanced Studies at Princeton, New Jersey, until his death. Gödel, like Tarski and Church, is best known as a logician. But his logical discoveries are of profound significance for parts of philosophy: the philosophy of logic and mathematics, epistemology, and (perhaps) the philosophy of mind. In addition, in later years Gödel concentrated on philosophical questions and made strikingly original suggestions as to their solution, including an improved version of Anselm's famous ontological argument for the existence of God, as elaborated by Leibniz.

Gödel's most famous discoveries are his two *incompleteness theorems* (Gödel 1931). Here we will give outlines of modernized version of his proofs. Suppose that the arithmetic of the natural numbers (0, 1, 2, and so on) is formulated as an axiomatic system. The language used is a precisely specified symbolic, or formalized, language with axioms stating the basic properties of the natural numbers and rules of inference stating which sentences may be correctly inferred from others. A sequence of sentences beginning with axioms and constructed by applying the rules of inference is said to be a *proof* of the last sentence in the sequence, which latter is said to be a *theorem* of the system.

Gödel observed that we can assign numbers (now called "gödel numbers") to the syntactical entities of the axiomatic system. That is, one can correlate numbers with symbols, with complex expressions, and even with sequences of expressions such as proofs. These numbers are assigned to symbols of the object language in the metalanguage. But these numbers are part of the subject matter of the object language theory itself.

This done, we have a sort of indirect way of talking about expressions and sequences of expressions of the formal theory within the theory itself. By talking about the gödel

numbers of expressions and sequences of expressions, we can simulate, or model, talk about the expressions and sequences.

Gödel then proved that if the formal system of arithmetic meets certain minimal conditions of adequacy, then the set of gödel numbers of the sentences provable in the system (the theorems of the system) can be defined within that system. The condition of minimal adequacy is that the system of arithmetic be capable of expressing certain functions of natural numbers. A function of natural numbers is just a correlation between numbers and numbers, or between pairs of numbers and numbers, or . . . and so on. Intuitively, the functions of natural numbers that must be expressible are those whose values for given arguments can be "effectively calculated": calculated by means of an algorithm or recipe, mechanically (and mindlessly) computed in a manner available to a computing machine.

Next, it can be proved that the set of gödel numbers of *true* sentences ("true" being defined in the manner of Tarski) is *not* definable in arithmetic. The proof uses an argument which parallels the reasoning of the liar antinomy (!) but which is logically unexceptionable. The conclusion is that such a system of arithmetic cannot contain or define its own truth predicate (as applied to gödel numbers as surrogates for sentences).

But if the set of gödel numbers of provable sentences *is* definable in arithmetic and the set of gödel numbers of true sentences (of arithmetic) is *not* definable in that system, then the two sets have to be different. Therefore, either some sentence provable in arithmetic is not true or some true sentence of arithmetic is not provable. We cannot accept the former, at least if we have chosen a system of axioms which we can see to be true of the natural numbers. We conclude that *some true sentence of arithmetic is not a theorem of arithmetic*. Any formal system of arithmetic meeting reasonable conditions of adequacy will be *incomplete*. This is essentially Gödel's first incompleteness theorem.

Gödel's actual proof did not proceed in the way we have described. Rather, he assumed not that the axioms of arithmetic are true (and that the rules of inference preserve truth) but only that arithmetic is *consistent*: it is not possible to derive an actual contradiction from the axioms using the rules of logic. (Really, he assumed that arithmetic is *omega-consistent*, a stronger assumption than consistency which we need not explain here.) Then Gödel showed how to construct, given that arithmetic is consistent, a particular sentence G which is such that neither G nor its negation –G ("not-G") is a theorem. The proof proceeds in such a way that we could, with sufficient patience and longevity, actually write down a true sentence of arithmetic which, if arithmetic is formally consistent, cannot be proved in arithmetic. And one can see that if the consistency of arithmetic is accepted, the sentence G is the true but unprovable one, as opposed to –G.

This sentence G involves just quite ordinary arithmetical concepts such as "plus" and "times" together with the usual logical concepts "and," "not," "some," "all," "equals," and so on. It is worth noticing that, contrary to various popular expositions, Gödel's original proof does not involve self-reference in any sense. The true but unprovable sentence G does not "say" that it, itself, is unprovable. It is a sentence entirely about natural numbers and their properties and relations. But it is a sentence that simulates such a self-referential sentence in the sense that it is true if and only if it is not provable.

Well, so what? It is natural to suggest that the axioms of arithmetic with which we began are just not adequate and that some new axioms must be added. It is as if Euclid's

geometry had been formulated without the Parallel Postulate. One would simply have to add it, or some equivalent. However, if you consider the details of Gödel's proof, it is evident that a system obtained by adding any finite number of axioms will still be subject to the proof – although the unprovable but true sentence will be different. Indeed, even adding infinitely many new axioms in any "effective" way does not evade the proof. We must conclude that nothing that would count as a formal system can contain all the truths of arithmetic. (If you think that to be a truth of arithmetic is to be provable in arithmetic, then this result will be quite difficult to comprehend. But one conclusion we can draw from these considerations is that this identification cannot be correct.)

This much is already quite startling. The goal of mathematicians since Euclid has been to specify certain basic truths of mathematics and to justify all others by deduction from these. Gödel proved that this goal is unattainable! No matter what formal (alias "axiomatic") system is proposed, there will be truths of arithmetic (the most basic part of mathematics) which the system cannot prove. Of course to show this with mathematical precision requires that we precisely define "axiomatic system" or "formal system," but this can be done in a way that is undeniably correct.

Gödel's second incompleteness theorem builds on the first. Using the technique of pseudo-self-reference mentioned above, one can find a sentence of any (minimally adequate) formal system of arithmetic which "says" that the system itself is consistent. Call this sentence "Consis." Now the proof of Gödel's first incompleteness theorem can be mimicked within arithmetic to produce a proof of the conditional sentence: "If Consis, then G." Suppose it were possible to prove within arithmetic the sentence Consis which mirrors the proposition that arithmetic is consistent. Then it would be possible to prove (by modus ponens) the Gödel sentence G. But we already know, from the first incompleteness theorem, that if arithmetic is consistent, G cannot be proved therein. If we suppose, as we are certainly entitled to do, that the theorems of arithmetic are true, then arithmetic is consistent. We conclude that the sentence which (in an indirect sense) expresses that arithmetic is consistent, cannot itself be proved in arithmetic. And, of course, the sentence "expressing" that arithmetic is *inconsistent* will not be a theorem either. It too, like the sentence G, is undecidable in arithmetic: can neither be proved nor refuted therein.

We have been at some pains to dispel the impression that Gödel's proofs literally involve self-reference. What then does, for example, the sentence "expressing" the consistency of arithmetic look like? Well, like the sentence G, it is written entirely in the language of arithmetic (involving "plus," "times," and "equals," for example). In fact it can be seen as expressing a certain mathematical claim about a *polynomial*. Let P(x,y) be a polynomial involving just the two indicated variables and integral coefficients. Then the statement "For every y, there is an x, such that $P(x,y) = 0$" may be true or false, depending on the details of the polynomial. Problems of this sort, as to whether or not such a statement is correct, are called "diophantine" problems. More generally, suppose we have n variables $x_1, x_2, \ldots, x_n$ and m variables $y_1, y_2, \ldots, y_m$, and a polynomial (with integral coefficients) involving these, say "$P(x_1, \ldots, x_n, y_1, \ldots, y_m)$." Then the statement "expressing" the consistency of arithmetic is of the form:

For every $y_1, \ldots, y_m$, there are $x_1, \ldots, x_n$, such that $P(x_1, \ldots, x_n, y_1, \ldots, y_m) = 0$.

This is a purely arithmetical statement and, one would have supposed, a claim that arithmetical techniques should be able to settle, yea or nay. But they cannot. (The details of these proofs are rather difficult to grasp. The best introduction to them is Smullyan 1957. For precise details, Boolos and Jeffrey 1989 is excellent.)

The philosophical import of these results is more controversial. We have already observed that a certain mathematical program is thereby proved impossible: that of deducing all of mathematics from an axiomatic basis. If we think of metaphysics as including all necessary truths, mathematics not excluded, then the goal, perhaps most closely associated with Spinoza, of an axiomatic development of metaphysics is thus also proved impossible to achieve.

Gödel's own philosophical speculations on the import of his theorems are most clearly articulated in his *Collected Works* (1995: 304–23). Let us confine our attention to arithmetic and speak of the true sentences thereof as "objective mathematics." Human beings, using our presently accepted arithmetical assumptions, can certainly prove some of these sentences. Gödel calls the mathematical truths that human beings are capable of demonstrating "subjective mathematics" (perhaps not the best choice of terminology since these are all objectively true and indeed knowable to be such). Some of these may require axioms about the numbers which are presently unknown, but which can in principle be seen to be evident by human mathematicians.

Now consider again Gödel's second incompleteness theorem. The theorems of a formal system of arithmetic comprise a set of sentences that could be mechanically generated, one after the other. According to Gödel's second incompleteness theorem, any such system will be unable to prove (generate) the arithmetical sentence that "expresses" its own consistency. Recall that this was a certain sentence expressing a diophantine arithmetical problem. Gödel draws the following conclusion from this:

> Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified. (1995: 310)

To say that the evident axioms (and rules of inference) can be "comprised in a finite rule" is equivalent to the possibility of the formulation of arithmetic as a whole as a formal system. And this latter amounts to the possibility of being generated in a machine-like fashion. To say that a diophantine problem is absolutely unsolvable means that neither the statement nor its negation will ever be a theorem of subjective mathematics. There is no doubt that this conclusion actually does follow from Gödel's proof, together with the mentioned analysis of a finite mechanical procedure. Gödel himself thinks, and argues, that the second option is incorrect – there are no absolutely unsolvable arithmetical problems – but his arguments for this conclusion are not airtight and may be reasonably doubted.

Gödel also had an improvement on Anselm's ontological argument for the existence of God, especially as it was developed by Leibniz. Leibniz observed that the one

thing that needs to be proved to complete Anselm's proof is that the existence of God is *possible*. A number of commentators on the argument, including Kant, have observed that all that is really established by Anselm is that *if* God exists, then He necessarily exists. Now, given the nature of the proof, we may further conclude that this conditional itself is necessary. So, we may conclude further, by a standard principle of modal logic, that if it is *possible* that God exists, then it is *possible* that it is necessary that He exists. (The standard principle of modal logic in question is this: if p necessarily implies q, then if p is possible, then q is possible.) If we further suppose that we can somehow prove that it is possible that God exists (given a certain definition of "God"), then it follows that it is necessary that it is possible that God exists. But according to one plausible system of modal logic (standardly called "S5"), it then follows that it is necessary that God exists!

But can we prove that it is possible that God exists? Gödel thought that we can and that a version of the ontological argument is then cogent. His argument for this, which bears some resemblance to Leibniz's argument for the same conclusion, uses the idea of a *positive* property. Gödel doesn't really say very clearly what this conception involves, but he remarks that it has two possible interpretations, "positive in the moral-aesthetic sense" and positive in the sense of involving only "pure attribution." A being is defined to be *God-like* if it has every positive property. Then a property is defined to be an *essence* of an entity x if x has that property and it entails every other property that x has. An entity *necessarily exists*, by definition, if every essence of it is necessarily exemplified. Finally, Gödel assumes the following "axioms" about these concepts:

1 A property is positive if and only if its negation is not positive.
2 Any property entailed by a positive property is itself positive.
3 The property of being God-like is positive.
4 If a property is positive, it is necessarily positive.
5 The property of necessarily existing is positive.

From these it follows that it is possible that a God-like being exists and, by essentially the argument explained earlier, that such a being therefore exists, and indeed necessarily so. There are some problems with the argument, not the least of which is the obscurity of the notion of a positive property. For an able discussion of the argument and its alleged defects, see Adams 1970.

Gödel also thought that Einstein's Theory of Relativity has implications for idealism, in particular that it supports some ideas of Immanuel Kant. He argues that there is considerable reason to believe that "time is unreal" (1951: 555–62). Essentially the argument is this: If time and change are something real, then there must be such a thing as an objective and absolute lapse of time. But the Theory of Relativity, a well-confirmed scientific theory, seems to deny that there is such an objective lapse. Gödel considers various objections and explains the relevance of a technical contribution of his to that theory. That work, by the way, seems to imply the possibility of "time travel"!

In sum, we may say that Gödel's main work in logic is of profound philosophical significance and that his other philosophical work certainly deserves further careful study.

Bibliography

Works by Church, Gödel, and Tarski

Church, A., 1950: "On Carnap's Analysis of Statements of Assertion and Belief," *Analysis* 10, pp. 97–9.
1951: "A Formulation of the Logic of Sense and Denotation," in *Structure, Method and Meaning*, ed. P. Henle, H. M. Kallen, and S. K. Langer, New York: Liberal Arts Press.
1956: *Introduction to Mathematical Logic*, vol. I, Princeton, NJ: Princeton University Press.
1973: "Outline of a Revised Formulation of the Logic of Sense and Denotation (Part I)," *Noûs* 7, pp. 24–33.
1974: "Outline of a Revised Formulation of the Logic of Sense and Denotation (Part II)," *Noûs* 8, pp. 135–56.
Gödel, K., 1931: "Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I," *Monatshefte für Mathematik und Physik* 38, pp. 173–98.
1951: "A remark about the relationship between relativity theory and idealistic philosophy," in *Albert Einstein: Philosopher-Scientist*, ed. P. A. Schilpp, New York: Tudor Publishing Company, pp. 555–62.
1986: *Collected Works*, vol. I, ed. S. Feferman, J. Dawson, and S. Kleene, New York: Oxford University Press.
1995: *Collected Works*, vol. III, ed. S. Feferman and J. Dawson, New York: Oxford University Press.
Tarski, A., 1944: "The Semantic Conception of Truth and the Foundations of Semantics," *Philosophy and Phenomenological Research* 4, pp. 341–76.
1956a: "On Definable Sets of Real Numbers," in J. H. Woodger, *Logic Semantics, Metamathematics: Papers from 1923 to 1938 by Alfred Tarski*, Oxford: Clarendon Press, pp. 110–42.
1956b: "On the Concept of Logical Consequence," in Woodger, pp. 409–20.
1956c: "The Concept of Truth in Formalized Languages," in Woodger, pp. 152–278.

Works by other authors

Adams, R. M. (1995) "Introductory note to *1970," in *Kurt Gödel: Collected Works*, vol. III, ed. S. Feferman, et al., Oxford: Oxford University Press, pp. 388–402.
Anderson, C. A. (1998) "Alonzo Church's contributions to philosophy and intensional logic," *Bulletin of Symbolic Logic* 4, pp. 129–71.
Boolos, G. S. and Jeffrey, R. C. (1989) *Computability and Logic*, 3rd edn., Cambridge: Cambridge University Press.
Smullyan, R. (1957) "Languages in which self reference is possible," *Journal of Symbolic Logic* 22, pp. 55–67.
Woodger, J. H. (trans.) (1956) *Logic Semantics, Metamathematics: Papers from 1923 to 1938 by Alfred Tarski*, Oxford: Clarendon Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

10

F. P. Ramsey (1903–1930)

BRAD ARMENDT

Frank Plumpton Ramsey made lasting contributions to philosophy, logic, mathematics, and economics in an astonishingly short period. He flourished during the 1920s at Cambridge University, and he interacted with the many notable figures there, including Russell, Moore, Keynes, and Wittgenstein. He was by no means a minor figure among this group. His work makes it clear, in fact, that he was at least their intellectual equal, and perhaps more, a judgment quite consistent with opinions expressed by his contemporaries.

Ramsey held an appointment in mathematics, but his main mathematical interests were in its foundations. He published just one piece of real mathematical work, a nine-page first section of his investigation of a decision procedure for a special case of first-order predicate logic ("On a Problem of Formal Logic," in Ramsey 1931). What he presented there as a preliminary tool has since been recognized as an important result – Ramsey's Theorem – and is the origin of a now thriving branch of mathematics, Ramsey Theory. He also published two papers in economics, one on taxation, and another on saving, that have also come to be regarded as important pioneering contributions to their subjects (in Ramsey 1978).

Ramsey had wide philosophical interests. He criticized and revised the logical system of *Principia Mathematica*, simplified its theory of types, and distinguished between the logical and semantic paradoxes. He gave a proto-functionalist account of belief, together with a redundancy theory of truth. He developed now influential accounts of partial belief, reasonable belief, probability, and knowledge. He developed an account of causality closely related to Hume's that has strongly influenced important contemporary accounts of laws of nature, and he made a proposal for representing the content of scientific theories via what are now known as Ramsey sentences. This is a distinguished collection of original contributions to philosophy; it is astonishing in light of the tragic fact that Ramsey died shortly before his twenty-seventh birthday, in January 1930. The circumstances of Ramsey's short life, and the time it took for many of his various projects to influence others, put us in an unusual position in studying and assessing his work. Ramsey was extremely productive, and his work always contains remarkable insight and originality. But his time was short, and probably none of his efforts approach the refinement of the treatments that we can imagine he would have produced, had he lived longer.

Foundations of mathematics

Though he later became a convert to finitism, Ramsey's well-known work on the foundations of mathematics explores and defends logicism, the view that mathematics is part of logic. Ramsey adopted the logic of Wittgenstein's *Tractatus*[1] and used its accounts of propositions and tautology to criticize and improve on the system of *Principia Mathematica* (*PM*). Russell and Whitehead used the ramified theory of types to avoid a variety of potential contradictions, including the paradox found by Russell that undermined the previous logicist theory of Frege (see FREGE and RUSSELL). *PM*'s ramified type theory involved a double hierarchy of (1) types of classes and propositional functions, and (2) orders of propositional functions within each type. Both hierarchies were motivated by, and according to Ramsey sloppily deduced from, a vicious circle principle: its idea is that classes may not include themselves as members, nor can propositional functions meaningfully apply to, or quantify over, themselves. With its arrangement of classes into types according to their membership, the system of *PM* rejected the paradoxical set of all sets not members of themselves. With its insistence that meaningful propositional functions be defined to apply only to functions of lower order, the system avoided further contradictions, such as Richard's and Grelling's paradoxes.

Ramsey criticized *PM* for recognizing only an impoverished range of classes, namely those definable through its comprehension axiom, but he had no objection to the type hierarchy it imposed on classes. He thought the ramified hierarchy of orders among functions of a given type was flawed in two ways, however. It missed the real nature of the paradoxes it sought to defuse, and it forced into the system of *PM* a nonlogical axiom, thereby undermining the logicist project. The axiom in question was the axiom of reducibility, asserting that for any propositional function of higher order there is an extensionally equivalent function of *lowest* order. With ramified types the axiom is needed, for instance, to guarantee that upper bounds of sets of real numbers will themselves be real numbers, rather than distinct entities with defining characteristics having a higher order than do the defining characteristics of the reals. Without the axiom *PM* could not capture important parts of mathematics (calculus and analysis, for example), but with it, Ramsey argued, *PM* made use of an axiom lacking logical necessity, and so failed its attempt to ground mathematics in logic alone.

Ramsey cites Peano for noticing that Richard's paradox is linguistic rather than mathematical, but he is the one remembered for drawing a general distinction between logical and semantic paradoxes. The contradiction threatened by Russell's set would occur within a mathematical system, but not so with many other paradoxes. The liar paradox, Richard's paradox and Grelling's paradox (which Ramsey attributes to Weyl) "occur not in mathematics, but in thinking about mathematics . . . [they] could not be constructed without introducing the relation of words to their meaning or some equivalent" (1990: 184, 200). *PM* made the mistake of treating the semantic paradoxes as logical ones, bringing in ramified type theory, and with it, the reducibility axiom. Ramsey reconceived propositions and logical necessity along the lines of the *Tractatus*, and he provided a revised, simpler theory of types, now applying to the various symbolic *expressions* of propositions, for which the axiom of reducibility

was not needed. "For me propositions in themselves have no orders; they are just different truth-functions of atomic propositions – a definite totality, depending only on what atomic propositions there are. Orders and illegitimate totalities only come in with the symbols we use to symbolize the facts in variously complicated ways" (1990: 211–12).

The logicist project was doomed by work done after Ramsey's death, but it seems Ramsey himself abandoned it before then. He had joked of preserving mathematics from the "Bolshevik menace of Brouwer and Weyl," but he later made extensive notes on intuitionist mathematics, and Braithwaite reports that he was converted to that view near the end of his life (1931: xii, 1990: 219, 1991a: 197–220).

Belief and truth

What is truth? For Ramsey, who takes truth to be first a property of beliefs and judgments, and only derivatively a property of sentences, this becomes the question, What is it for a belief to be true? His answer is that a belief is true when it is a belief that *p*, and *p*. He regarded this as entirely obvious; the difficult part in analyzing the truth of belief is not with the concept of truth, but with the analysis of what it is to believe that *p*. To say that it is true that the earth is round amounts to saying that if anyone were to believe that the earth is round, their belief would be true. Which is to say "no more than that the earth has the quality you think it has when you think it is round, i.e. that the earth is round" (1990: 38–9, 1991b: 7–13). To say, "Everything he believes is true" amounts to no more than "For all *p*, if he believes *p*, then *p*." The latter sounds odd in ordinary discourse; our grammatical habits demand a verb and push us toward tacking an "is true" on to the end of the latter sentence, which then would hardly be enlightening. But the verb and the clause are unnecessary; a verb is already present in any of the beliefs he may have.

Ramsey regarded his account of truth as a qualified correspondence theory, but it has come to be known as a redundancy theory. As he made clear, if in order to give an account of what it is to believe that *p*, he were forced to rely on the concept of truth, then not much progress would have been made. So what accounts for the contents of beliefs, judgments, and assertions? Ramsey sketched what he regarded as a pragmatist account, influenced by Russell and Peirce, that looks to the causal properties of our mental states. The primary target of his account is occurrent linguistically expressed belief, consciously asserted or denied; the contents of dispositional beliefs are derivative from the contents of occurrent thoughts.[2] The relevant mental states are silent or spoken linguistic utterances accompanied by feelings of assent or denial, and their contents are given by causal properties they bear to other mental states and to the world. Beyond describing the general picture, Ramsey devoted most of his attention to explaining how such causal properties exhibit patterns corresponding to the logical structures of the propositional contents born by the states. The key point is that his account of belief generates his account of meaning, rather than the other way around. "The essence of pragmatism I take to be this, that the meaning of a sentence is to be defined by reference to the actions to which asserting it would lead, or, more vaguely still, by its possible causes and effects" (1990: 51).

Reasonable belief, probability, and knowledge

"Truth and Probability" was written in 1926 and published posthumously (1931).[3] It is best known for its treatment of partial belief and subjective probability. Beyond this, however, Ramsey presented a view of logic construed as the science of rational thought, which he divided into two parts, the logic of consistency and the logic of truth.[4]

At the core of the views now known as probabilism, Bayesianism (in epistemology or decision theory), and subjectivism (about probability) is the idea of *partial belief*, and there is still no better introduction to it than the third section of "Truth and Probability." We hold some of our beliefs more strongly than we hold others. What is it that varies among the beliefs held with different strengths? Different beliefs will have different contents, of course, and sometimes the content of a belief may be about the strength of a belief, as when I believe that you strongly believe that *p*, or doubt that *q*. I might have a view about the strength of my own belief, but that view (a second-order belief, we might say) is distinct from its subject (my belief that *p*, with whatever strength it has). So if it is not in the content of a belief, where and what is the strength of a belief? And if we can answer that, can we go on to make systematic sense of the degrees of strength that beliefs can and do have? Perhaps we can do little better than to say that on different occasions we are certain, or confident, or think that maybe . . . , and so on. But Ramsey's answer to the first problem (What is strength of belief?) illuminates the *point* of distinguishing different degrees of strength, and it makes room for a range of strengths richer and more systematic than is the range indicated by our ordinary reports.

The essential idea is that the most fruitful way to think about the degrees of strength to which we hold beliefs (degrees of belief, for short), is to attend to the ways beliefs guide us in our choosing and acting. A degree of belief is a causal property of it, "which we can express vaguely as the extent to which we are prepared to act on it" (1990: 65). Ramsey proposed, first, that there is a univocal way in which a proposition *p* enters into a person's deliberations, so that it makes sense to speak of *the* degree to which she holds *p*.[5] If we are willing to accept that, then we can look to a person's deliberations and potential choices for indications of *p*'s influence on them, i.e. for indications of her degree of belief in *p*. In doing this, we are exploring a measurement problem, and Ramsey was well aware of both the theoretical difficulties and the practical complications that accompany a solution to it. On both scores, he drew an analogy to measurement in physical science. On the theoretical side, just as the length of a time interval between two events depends in relativistic physics on exactly how it is measured, so may the influence of a belief on choice. So just as when we use the idea of time intervals, when we use the idea of degrees of belief we should keep track of how we propose to measure them, but "for many purposes we can assume that the alternative ways of measuring it lead to [approximately] the same result" (1990: 63; see also p. 68). On the practical side, the intertwining of different physical influences and the disturbances introduced by measurement processes do not undermine all our attempts to understand and quantify physical phenomena. So it is at least not obvious that similar practical complications in studying belief and choice will defeat our attempts to do so.

Ramsey first gave a familiar, if piecemeal, account. As he put it, "The old-established way of measuring a person's belief is to propose a bet, and see what are the lowest odds

which he will accept. This method I regard as fundamentally sound; [though inexact and not general]" (1990: 68). The odds one will offer on a bet (ratios of its potential payoffs) indicate one's degree of belief. If I am willing to offer high odds as I defend *p*, my degree of belief in *p* is high; if I offer only low odds (demand high odds from my opponent) my degree of belief is low. A *conditional* degree of belief in *p*, given that *q*, is indicated by the odds I would place on a wager on *p* which only pays off in the circumstance that *q* is true. Of course we undertake wagers infrequently compared to the frequency with which we have beliefs, and the betting scenario is quite artificial as a model for the variety of choices we make. Beliefs and their strengths are dispositions, though, and the measurements proposed via wagering are *indicators* of them, not the degrees of belief themselves. And in a wider sense, Ramsey said, all our lives we are betting: "Whenever we go to the station we are betting that a train will really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home" (1990: 79).

What degrees of belief ought we to have? The logic of consistency requires that degrees of belief obey the rules of probability. Apropos of the betting scenario, Ramsey stated – and seems to have understood better than most – what has become known as the Dutch book argument. Degrees of belief that violate probability would guide a person toward betting arrangements guaranteed to yield a loss (a certain loss, according to his own values) and Ramsey took this to be an indication of inconsistency in the partial beliefs.

The most remarkable part of Ramsey's treatment of partial belief is his generalization of the betting scenario used so far. Generalize the idea of a bet on *p* to a gamble having the form, α if *p*, β if not, where α and β represent states of the world bearing value and obtaining according to whether or not *p* does. Some gambles will be favored over others, depending upon the values of α and β, and on *p*. Ramsey showed that if a person's preferences among a rich set of these gambles is well arranged, according to stated principles (transitivity, for example, is one), then there are non-arbitrary measurements of all the values and of the probabilities of all the propositions *p*. These measurements are attributable to that person's values and beliefs, and can be taken to be the values and degrees of belief that guide his choices. This result is an early forerunner of similar demonstrations given by many later economists and philosophers, results that are usually taken as foundations for utility theory or decision theory. Ramsey emphasized the importance of the result for an account of partial *belief*, though this theory yields both. Important later theories that follow him in this are Savage's and Jeffrey's.[6]

Beyond the dictates of consistency, what degrees of belief are reasonable? With repeated acknowledgment of Peirce's influence, Ramsey conceived of the logic of truth along pragmatist lines. The best approach is to ask about the reasonability of the *habits* by which we arrive at and hold our beliefs. *Always fully believe the truth* is not bad advice, but it is not a very useful recommendation either. Nor is it useful as a standard for the sort of general habits open to humans, habits that so often yield partial beliefs rather than certainties. A more appropriate standard judges the habits according to how closely their partial beliefs correspond to the rate at which the beliefs are true. That is, the habit should yield a partial belief whose strength corresponds to the frequency with which relevantly similar beliefs are true. Ramsey used an illustration involving a belief

about the wholesomeness of yellow toadstools; we can use the beliefs of weather forecasters. A forecaster does well when it rains 70 percent of the time in which her degree of belief is 7/10. Ramsey's work on the *consistency* of partial belief is well known. It is worth emphasizing that he attached as much importance to this second standard calling for alignment of degrees of belief with frequencies of truth (and, further, that he was aware of the complexities of developing it, e.g., in identifying habits and specifying what classes of cases are relevant). To return to the action-guiding nature of partial belief:

> [B]elief of degree m/n is the sort of belief which leads to the action which would be best if repeated n times in m of which the proposition is true. . . . It is this connection between partial belief and frequency, which enables us to use the calculus of frequencies as a calculus of consistent partial belief. And in a sense we may say that the two interpretations are the objective and subjective aspects of the same inner meaning, just as formal logic can be interpreted objectively as a body of tautology and subjectively as the laws of consistent thought. (1990: 84)

Ramsey said less about knowledge than about reasonable belief. He regarded knowledge as true, certain belief produced by a reliable process, and so it appears he considered it the extreme case of fully held true belief, backed up by a process that tends to produce such beliefs. This is a natural extension of his suggestion for evaluating our habits of belief, especially if the required reliability matches the strength of the belief (certainty). Though his most straightforward and explicit statements demand certainty, they are accompanied by discussions of fallibilism and by further remarks that soften the demand to near-certainty, practical certainty, or conviction "just a minute fraction short" of certainty (1990: 110–11, 1991b: 62–4). He agrees with Russell's view that "all our knowledge is infected with some degree of doubt," and in a paragraph on Moore's paradox and the paradox of the preface, he shifts from talk of certainty to talk of being *nearly* certain. It is likely a mistake to overanalyze his brief unpublished remarks, and it is difficult to determine how closely Ramsey's account of knowledge is tied to his account of reasonable belief. It is clear, though, that he endorsed a reliable-process account of knowledge, and he is remembered by contemporary epistemologists for doing so.

Laws, causality, and theories

In the last two years of his life Ramsey worked seriously on causality, laws of nature, and the formal structure of scientific theories. The several papers on these topics are not so finished as his earlier work, and in places indicate rapidly evolving views. This is also the period in which he was at work on his book on truth and moving away from logicism to an intuitionist view of mathematics.

Ramsey's view of causality was not very distant from Hume's. In his brief 1928 paper on law and causality (1990: 142–3), he suggests that the difference between universals of law and universals of fact (between lawlike and accidental generalizations, we might say) lies in their distinct roles in our system of knowledge. If we knew everything and organized our knowledge in a deductive system that strove for simplicity, the

general axioms of the system would be the fundamental laws of nature. They and the generalizations derivable from them without reference to facts of existence are the "statements of causal implication." And though really we do not know everything, we do tend to organize our knowledge in a deductive system, regard its axioms as laws, and regard as undiscovered laws the future axioms we expect to arise as we learn more. Ramsey soon revised this view, but its influence persists in the work of more recent philosophers, notably in David Lewis's best-system account of laws[7] (see LEWIS).

Ramsey's revised treatment is in the 1929 paper, "General Propositions and Causality." The view there is that causal laws do not get their force by being simple fundamental generalizations in an axiomatic summary of our knowledge. Their causal force lies in our trusting them as guides in our inferences about particular events. Causal generalizations "are not judgments but rules for judging 'If I meet a ϕ, I shall regard it as a ψ.' This cannot be *negated* but it can be *disagreed* with by one who does not adopt it." An assertion of a causal law is an assertion not of a proposition, but of a formula from which we derive propositions about particular events. Its causal character lies in the temporal ordering of the events about which it licenses our judgments (ψ does not precede ϕ). The special importance we attach to judgments with that ordering is traceable to the importance of forward-looking judgments in our thinking about the influence our actions may have on the world.[8] In the course of discussing conditionals in this paper, Ramsey suggests that the acceptability of a conditional goes by the acceptability of its consequent after the antecedent is hypothetically added to one's beliefs:

> If two people are arguing "If *p* will *q*?" and are both in doubt as to *p*, they are adding *p* hypothetically to their stock of knowledge and arguing on that basis about *q*; . . . We can say that they are fixing their degrees of belief in *q* given *p* . . . (1990: 155)

In contemporary work on conditionals this idea has become widely known as the *Ramsey test* for the acceptability of a conditional.

In another 1929 paper, Ramsey addresses the formal structure of scientific theories ("Theories," in Ramsey 1990). He is particularly interested in the question of the content of theoretical assertions, and how such content is related to the observational assertions on which the theory is built. One idea for demonstrating the dispensability of theoretical terms is to show that they are explicitly definable in terms of the observational assertions, and further, that the definitions can be inverted so that anything we express in theoretical language can also be expressed without recourse to the theoretical terms. Ramsey works all this out for a very simple example; even for that example, the results are complex and extremely cumbersome. Worse, as he points out, the method of explicit definitions creates the problem that additional data not seriously inconsistent with the theory will nevertheless falsify the theory unless adjustments are made to the meanings of its terms. Is there another alternative? Ramsey offers one. Conjoin all the sentences of the (first-order) theory, replace the occurrences of each distinct theoretical term with a second-order variable, and introduce for each distinct variable a second-order existential quantifier that binds its occurrences. The result is now known as the *Ramsey sentence* of the theory. It contains only observational terms, is entailed by the first-order theory, and it entails the

same particular observation sentences as the first-order theory. This device has since been used by Hempel, Carnap, and many others in treatments of the content and meaning of theories, whether the concern is, as above, with the observational content of scientific theories or with, for example, the relation between mental and neurophysical theories.

Notes

1 L. Wittgenstein, *Tractatus Logico-Philosophicus*, London: Routledge, 1922. Ramsey contributed to the 1922 English translation and made criticisms that led to changes in the 1933 second edition. Wittgenstein mentions Ramsey's influence in the preface to his *Philosophical Investigations*.
2 Dispositional linguistic belief, that is; Ramsey also entertained the idea that a chicken has beliefs, to be understood as relations between its potential behavior and circumstances in which the behavior is appropriate. This resembles the analysis he gives of dispositional strengths of belief (see below); Ramsey 1990: 40.
3 Several later notes expressing afterthoughts and further ideas are in Ramsey 1931, 1991a.
4 The organization of "Truth and Probability" was the starting point for a planned book. Much of the unfinished manuscript for the book has been published as Ramsey 1991b.
5 Notice that the idea of a single degree of belief is for a person, at a time, among a collection of other beliefs. A degree of belief may very well change for a variety of reasons, including observations, the passage of time, or changes in other beliefs.
6 L. J. Savage, *The Foundations of Statistics* [1954], 2nd edn., New York: Dover, 1972; R. C. Jeffrey, *The Logic of Decision* [1965], 2nd edn., Chicago: University of Chicago Press, 1983.
7 D. Lewis, *Philosophical Papers*, vol. II, Oxford: Oxford University Press, 1986. Lewis acknowledges "following the lead of (a short temporal segment of) Ramsey," p. xi.
8 Ramsey calls these rules for judging *variable hypotheticals*, of which causal laws are an important kind. Ramsey 1990: 149, 157–9.

Bibliography

Ramsey's best-known and most influential work is found in the collections of 1931, 1978, and 1990; their philosophical contents are very similar. His surviving papers are in the Ramsey Collection housed at the University of Pittsburgh's Hillman Library. A fascinating group of these papers, almost all otherwise unpublished, appear in the collection dated 1991a. A substantial manuscript on truth, also from the Ramsey Collection, was published in 1991b. The contents of Mellor (1980) are more inspired by Ramsey than centered on his work. Bibliographies of Ramsey's work appear in Ramsey 1931, 1978, 1990, and in Sahlin 1990; all of the books listed below contain interesting introductions or prefaces.

Works by Ramsey

1931: *The Foundations of Mathematics and Other Logical Essays*, ed. R. B. Braithwaite, London: Routledge and Kegan Paul.
1978: *Foundations*, ed. D. H. Mellor, Atlantic Highlands, NJ: Humanities Press.
1990: *Philosophical Papers*, ed. D. H. Mellor, Cambridge: Cambridge University Press.
1991a: *Notes on Philosophy, Probability, and Mathematics*, ed. M. C. Galavotti, Naples: Bibliopolis.
1991b: *On Truth*, ed. N. Rescher and U. Majer, Dordrecht: Kluwer.

Works by other authors

Mellor, D. H. (ed.) (1980) *Prospects for Pragmatism: Essays in Memory of F. P. Ramsey*, Cambridge: Cambridge University Press.

Sahlin, N. (1990) *The Philosophy of F. P. Ramsey*, Cambridge: Cambridge University Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

11

Carl G. Hempel (1905–1997)

PHILIP KITCHER

Introduction

Carl Gustav Hempel was one of a group of philosophers from Central Europe who emigrated to the United States in the 1930s and who profoundly modified the character of American philosophy. Together with Rudolf Carnap, Ernest Nagel, and Hans Reichenbach, Hempel was central to the transition from logical positivism to logical empiricism. His writings not only set the agenda for philosophy of science in the middle decades of the twentieth century but also continue to shape this important field of philosophy.

Educated in Berlin, Hempel was influenced by early twentieth-century attempts to apply the concepts and techniques of mathematical logic to the empirical sciences, pioneered by Reichenbach in Germany and Carnap and his co-workers in Vienna. The Vienna Circle had hoped to diagnose most traditional philosophical discussions as treatments of pseudo-problems, by formulating and applying a precise criterion of cognitive significance. The envisaged criterion was intended to pick out, as meaningful, statements of logic and mathematics conceived as analytic truths, and those non-analytic statements that admit of empirical test. The sciences, paradigmatically the physical sciences, were to count as meaningful because they satisfied the latter condition, whereas the statements of traditional philosophy were to be exposed as neither analytic truths nor susceptible to empirical test, and consequently devoid of cognitive significance.

During the 1930s, logical positivists made successive attempts to make the criterion of cognitive significance sufficiently sharp to perform the planned surgery. Hempel participated actively in these discussions, examining various proposals and identifying difficulties. His efforts culminated in synthetic essays that argued for the impossibility of the positivist project and drew the blueprint for logical empiricism, the philosophical structure within which he would build substantive positions.

Cognitive significance

Hempel never wavered in his commitment to the idea that "the general intent of the empiricist criterion of meaning is basically sound" (1965: 102), that is, that meaningful empirical statements are those that admit of experiential test. He believed,

however, that attempts to state precise logical criteria of empirical significance encounter systematic difficulties. If we insist that empirical statements are those that admit of conclusive verification, then the criterion will debar universal statements (including the laws of the various sciences); to propose that empirical statements must allow conclusive falsification is equally hopeless, since this would deny meaningfulness to existential claims. Nor can we propose to formulate a criterion of meaningfulness by requiring definability of all terms in some language whose nonlogical vocabulary can be learned in application to observational entities and properties, for the special terms that play so fruitful a role in the physical sciences, expressions like "state function" and "covalent bond" cannot be given explicit definitions. Hempel concludes that the requirement of meaningfulness must imitate the way in which scientists extend the resources of everyday language, to wit by systematically connecting sentences in which unfamiliar expressions appear with sentences whose vocabulary is unproblematic.

My reconstruction condenses a wealth of subtle points that Hempel articulates with characteristic lucidity, but it brings out two important aspects of the critique of formal criteria of cognitive significance. First, Hempel is not simply concerned with a philosophical enterprise that intends to transcend what are taken to be sterile disputes, but, in addition, to provide an assessment of movements in the natural and social sciences that call for empirical conditions for the application of bits of theoretical vocabulary. He offers a definitive rebuttal to the operationalist demand that newly-introduced terms must be associated with an experiential procedure for applying them (see Hempel 1965: ch. 5). Second, that rebuttal depends on taking the unit of appraisal to be "sentences forming a theoretical system" (1965: 117). Explicitly acknowledging that this makes cognitive significance, "a matter of degree" (ibid.), he offers four main ways in which theoretical systems should be evaluated: by scrutinizing the connections among theoretical terms and the connections to statements couched in observational vocabulary; by considering the explanatory and predictive power of the system; by appraising the simplicity of the system; and by assessing the degree of confirmation by empirical evidence.

For the positivist eager to prick the pretensions of traditional metaphysics or the operationalist concerned that some area of science should be rigorously developed, this concluding catalogue must come as a disappointment. In place of a clear criterion that can put an end to disputes, it appears that these fuzzy virtues of theoretical systems will be hard to identify and that we are fated to continue the wrangles of the past. Hempel's reply would be twofold. First, he would point out that the weapons for which positivists and operationalists yearn are simply not to be had. Second, he would insist that his list of modes of appraisal may be the end of a reflection on the fate of positivism, but that it is only the beginning of a serious philosophy of science. The task for logical empiricism is to say, as precisely as possible, what kinds of linguistic connections are present in virtuous theoretical systems, what makes for explanatory and predictive power, what counts as simplicity, and how empirical statements are confirmed by the results of experiential tests.

For Hempel, then, there are four main problems of the philosophy of science. The last problem, the issue of empirical confirmation, is also central to the theory of knowledge, and, since ontological questions depend on obtaining a clear view of the

kinds of entities to which the sciences commit us, clarification of theoretical structure is pertinent to discussions in metaphysics. Throughout his career, Hempel made major contributions to three of the four problems he highlighted: he offered a theory of qualitative confirmation, provided magisterial discussion of the fruitful use of theoretical vocabulary, and delineated an account of scientific explanation against which all subsequent treatments of this topic must be measured. Other logical empiricists, particularly Reichenbach (1938) and Nelson Goodman (1949), took up the problem of understanding simplicity, although this topic never achieved the same prominence within logical empiricism as the other issues.

A philosopher's life work rarely conforms to a neat structure, and Hempel's is no exception. During the course of his career, he wrote important essays on the character of mathematics and the relations between mathematics and the natural sciences (1945a, 1945b). He also retained a strong interest in the properties of systems of classification, from his early attempt to apply logical notions in taxonomy (Hempel and Oppenheim 1936) to various attempts to evaluate proposed classificatory schemes in psychiatry and in the social sciences (1965: chs 6 and 7). Despite the undeniable influence that these studies have had, Hempel's attacks on the problems of confirmation, theory-structure, and explanation are, I believe, his most enduring accomplishments. The following sections will consider them in order of ascending importance.

Qualitative confirmation

Suppose that h is a hypothesis in whose truth-value we are interested, and that e is a statement that reports the result of some empirical test. The general problem of confirmation is to understand the nature of the relation of evidential support between h and e. This general problem encompasses several specific questions: we might ask the *degree* to which e would support h (the quantitative problem); alternatively, we might inquire after the conditions under which e would support h at all (the qualitative problem); an intermediate question probes the conditions under which e supports h more than e' supports h' (the comparative problem; note that e might be the same as e' or h identical with h'). In his extensive investigations of inductive logic, Carnap focused on the quantitative problem, considering formal languages adequate for the formulation of fragments of science, and attempting to define, for a wide class of statements h and e within these languages, the degree to which e would confirm h. By contrast, Hempel took the qualitative problem to be more fundamental, and endeavored to specify conditions under which singular statements (conceived as ascribing properties to objects) would confirm, disconfirm, or be neutral to, a hypothesis (characteristically thought of as a lawlike generalization).

Hempel's treatment is noteworthy not just for his positive proposal but for his disclosure of interesting difficulties with apparently plausible ideas. Suppose that the hypothesis of interest is the generalization that all ravens are black, formalized as $(x)(Rx \supset Bx)$. It is very natural to believe that the generalization is supported by observing black ravens. So we might arrive at the general proposal of confirmation by instances: the hypothesis $(x)(Rx \supset Bx)$ is confirmed by any sentence $Ra\&Ba$. As Hempel points out, the manner in which we formulate a hypothesis should make no difference to the class of statements that confirm it. Thus, if h and h' are logically equivalent any

e that confirms h should confirm h', and conversely. By elementary logic, $(x)(Rx \supset Bx)$ is logically equivalent to $(x)(-Bx \supset -Rx)$. The thesis of instance confirmation now tells us that the latter is confirmed by $-Ba\&-Ra$, which, by the principle about logical equivalence, must also confirm $(x)(Rx \supset Bx)$. Reverting to our interpretation of the nonlogical vocabulary, we discover that the generalization "All ravens are black" is confirmed by statements that tell us that a particular object is neither black, nor a raven. Apparently, learning that the rightmost shoe in my closet is white would support an ornithological generalization!

The "paradox of the ravens" has inspired a large subsequent literature. Hempel's own diagnosis was that there is no genuine paradox, and that any sense of surprise stems from "misguided intuitions" (1965: 20). Whether or not this is so, there is no doubt that Hempel's further investigations disclose severe problems in natural conceptions. Many philosophers, and scientists reflecting on methodological issues, have accepted the fundamental idea of hypothetico-deductivism, to wit that hypotheses are confirmed when their consequences are found to be correct. The most straightforward way to formulate that idea is as the *Converse Consequence Condition*: if e is a consequence of h, then e confirms h. Unfortunately, that condition, coupled with a requirement that is hard to resist, generates the conclusion that any statement will confirm any hypothesis. The further requirement is the *Entailment Condition*: evidence statements confirm the logical consequences of the hypotheses they confirm. Given any statement e, e is a consequence of $h\&e$ (whatever h may be); so by the Converse Consequence Condition e confirms $h\&e$; h is a consequence of $h\&e$; hence by the Entailment Condition e confirms h.

Hempel's own account of qualitative confirmation avoided this difficulty by abandoning the Converse Consequence Condition. Instead, he proposed that direct confirmation results when an evidence statement entails a restricted version of the hypothesis, effectively what the hypothesis would say if there just existed the individuals mentioned in the evidence statement. More exactly, suppose that the evidence statement e contains the names of exactly the individuals $a_1, \ldots, a_n$; then e directly confirms $(x)(Rx \supset Bx)$ just in case e entails each of the statements $Ra_1 \supset Ba_1, \ldots, Ra_n \supset Ba_n$. The general notion of confirmation is obtained by proposing that e confirms h just in case h is entailed by a class of sentences each member of which is directly confirmed by e.

Hempel's approach is unable to account for the confirmation of sentences containing theoretical vocabulary by statements formulated in more basic terms. Clark Glymour attempted to extend the Hempelian approach to offer an account of qualitative confirmation (or of relevant evidence) that would address this difficulty (Glymour 1980), and his proposal has given rise to extensive subsequent discussion. Hempel himself became convinced that the general approach could not succeed, on the grounds that Nelson Goodman's "new riddle of induction" showed the inadequacy of any purely syntactical analysis of qualitative confirmation (see GOODMAN). Arguing that not all universal generalizations are supported by their instances, Goodman (1955) exposed the difficulties of distinguishing those generalizations that can be confirmed in this way from those that cannot.

Ironically, much contemporary thinking about confirmation diverges from Hempel's treatment at a very early stage. The most influential more recent proposal is

Bayesianism, a position that descends from Carnap's investigations in inductive logic and that attempts to understand how degrees of confirmation of hypotheses adjust in the light of evidence. For Bayesians, the problem of quantitative confirmation is primary and the solution to the qualitative problem is generated in a trivial fashion from a solution to the quantitative problem. To solve the latter, we need to be able to assess the value of the probability of the hypothesis given the evidence, $Pr(h|e)$; assuming that that can be done, we can say that e confirms h (or confirms h relative to background information B) just in case $Pr(h|e) > Pr(h)$ (or $Pr(h|e\&B) > Pr(h|B)$). Hempel's discussions of confirmation remain of interest not so much because of his positive proposal about qualitative confirmation as for his careful exposure of difficulties in intuitively attractive ways of thinking about confirmation, and for his recognition of constraints on any adequate solution.

Theories

Logical empiricism began with the conviction that the tools of logic developed by Frege, Russell, and their successors could be used to make clear and explicit the structure of scientific theories (see FREGE and RUSSELL), and, indeed, even in the 1920s, Reichenbach had offered an axiomatization of the special theory of relativity, intended to exhibit which parts of the theory were conventional stipulations and which made substantive empirical claims. Almost unselfconsciously, the logical empiricists took over the logician's conception of a theory as a deductively closed set of sentences in some suitable formal language. Recognizing that an important – and, from the empiricist viewpoint, problematic – feature of the theories in physics and chemistry that most impressed them was the presence of special vocabulary that resists explicit definition in observational terms, logical empiricists formulated a distinctive view about scientific theories. A scientific theory is a deductively closed set of statements in a first-order language whose nonlogical vocabulary divides into two subsets, the basic vocabulary (often understood as containing those terms whose application can be made on the basis of more or less direct observation) and the theoretical vocabulary (the remainder); the statements whose essential nonlogical vocabulary contains only theoretical terms are the theoretical postulates of the theory, while statements whose essential nonlogical vocabulary contains both theoretical and basic terms are the correspondence rules (a variety of other designations for this last class of statements appears in logical empiricist writings, but "correspondence rules" is the most popular locution); the function of the correspondence rules is to provide the theories with empirical content, and they (or a subset of them) are often conceived as providing an interpretation (or partial interpretation) of the theoretical vocabulary.

This conception of scientific theories became fully explicit in the writings of Carnap (1956), Nagel (1962), and Hempel (1958, 1965: ch. 8), and, perhaps as a residue of the concerns about cognitive significance, each of these authors sought ways to characterize those theoretical contexts in which the introduction of theoretical vocabulary served important scientific purposes (see CARNAP). Hempel's discussions revolved around a family of questions. To what extent can the correspondence rules be viewed as functioning as definitions? Is it problematic to concede that the correspondence rules only offer partial interpretations of the theoretical vocabulary, or should this be seen

as a symptom of the openendedness of scientific research? Can we eliminate the theoretical vocabulary without any scientific loss? Is it reasonable to treat the theoretical terms as components of a formal apparatus for making experiential predictions, or should we suppose that those terms refer to entities and properties that underlie the observable phenomena?

Hempel's discussion of these questions embodies a cautiously realistic attitude. He does not think that we can provide a full explicit definition of theoretical vocabulary in observational terms, not even for those relatively low-level parts of science that make use of dispositional concepts. Although he believes that correspondence rules are vehicles of partial interpretation, he suggests that we cannot neatly separate the parts of a theory that function to pin down the meanings of our terms from those that make genuinely empirical claims, a point in which he concurs with W. V. Quine's celebrated critique of the analytic/synthetic distinction (Quine 1953) (see QUINE). Partial interpretation, Hempel believes, has heuristic advantages, allowing us to introduce new correspondence rules as we extend the theory to cope with previously untreated phenomena. Further, to the extent that theoretical vocabulary is ineliminable, he holds that the evidence leading us to adopt the theory ought to incline us to accept its postulates as true and thus to treat its theoretical vocabulary as referring to entities beyond the reach of ordinary observation.

The crucial issue thus turns out to be whether or not there is a generally available method for eliminating theoretical vocabulary. Hempel approaches the problem by formulating a dilemma. Starting from the premise that the function of a theory is to "establish definite connections among observable phenomena" (1965: 186), he suggests that when such connections are established we do not need any detour through a theory, since the theory will imply a conditional statement, couched in the basic vocabulary, that asserts the connection. So, if the theoretical terms and principles serve their purpose, they can, in principle, be eliminated, and are thus unnecessary. If, on the other hand, the theoretical terms and principles do not establish the intended connections, they do not serve their purpose, and are consequently unnecessary. This is "the theoretician's dilemma" (1965: ch. 8).

Behind the dilemma stands a view of the use of theories in scientific practice. It is as though the scientist feeds some description of observables into the theoretical machinery, the gears turn, and the output is another statement about observables. The point may be to predict something (the output statement is one we didn't know before) or it may be to explain something (we already knew the output statement but didn't see why it was true). In either case, the theoretical statements function to license an inference from the input to the output, and thus must support the conditional statement, "If input then output." Why then can we not manage with the set of all such conditional statements corresponding to the transitions that the theory would license?

A first response is that the resultant set would be extremely unwieldy, that the theory as actually presented provides a concise way of representing a disparate class of consequences. As Hempel and his colleagues clearly saw, however, a result due to the logician William Craig constructs a recursive procedure for generating the class of consequences, without stepping outside the basic vocabulary. (It is interesting to reflect that the significance attributed to Craig's theorem revealed the hold that the logician's conception of theories as recursively axiomatizable continued to exert on logical empiricist

discussions of scientific theories.) In the end, Hempel's response to this and to kindred suggestions for eliminating theoretical vocabulary draws on the proposal that the task of a scientific theory is not only to achieve deductive systematization of observable phenomena but also to provide inductive systematization, and we have no reason to believe that any of the elimination procedures will satisfy this further constraint (1965: 214–15; note that in the case of the Craigian surrogate, Hempel is able to argue that the substitute will not achieve any inductive systematization).

It is at first sight ironic that the philosopher who contributed most to our understanding of scientific explanation overlooked a relatively obvious point about proposals for eliminating theoretical terms: even though they might be able to mimic the predictive successes of genuine theories, it seems that they would incur severe explanatory losses. Whatever set of brute empirical rules we might devise for predicting the outcomes of bringing substances together in various proportions would fail to deliver the explanatory benefits we obtain from embedding empirical generalizations within the theoretical treatment of shell-filling and of ionic and covalent bonds. On deeper reflection, however, we can see that appeal to explanatory power would have led Hempel into uncomfortable questions about the sufficiency of his preferred account of scientific explanation. For there is every reason to think that some of the procedures for eliminating theoretical terms would not just deliver singular conditional statements but generalizations from which such singular conditionals could be derived; because the generalizations would serve as the needed "covering laws" in Hempel's schemata for explanation, by the standards of the Hempelian account of explanation the explanatory loss would be indiscernible.

During the 1960s and 1970s, the account of theories favored by Hempel, Nagel, Reichenbach, and other logical empiricists acquired the name "the received view," and like most doctrines so designated came under vigorous attack (see Suppe 1970 for thorough analysis). One principal difficulty, recognized by Hilary Putnam, was that the account conflated two distinctions, the distinction between observational and theoretical *terms* and that between observable and unobservable *things*; as Putnam noted, some theoretical terms name observable things ("oscilloscope") and unobservable things can be picked out using observational terms ("people too little to see"). Putnam's observations, and the discussions of reference that he and others initiated (Kripke 1971, Putnam 1973) undercut the old concern that scientists are simply unable to specify the referents of their theoretical vocabularies.

A different line of objection attacked the idea that a theory is a linguistic item. Several authors (including Suppes 1967, van Fraassen 1980) drew inspiration from model theory rather than from the syntax of logical systems, proposing that theories are to be identified with families of models. Their critiques initiated a debate, as yet unresolved, about the correct analysis of the notion of a scientific theory. It is perhaps worth recalling that both the newer "semantic conception of theories" and the older "syntactic account" (or "received view") are philosophical reconstructions of the practices of scientists, and that the standards of adequacy for a reconstruction depend on what purposes – philosophical or scientific – one intends to achieve. The question "What is the *real* structure of a scientific theory?" may simply be a bad question, and, depending on our aims, we may draw on one or another of the proposed accounts (or on something completely different). To the extent that the syntactic conception con-

tinues to be valuable in such enterprises, we can expect that Hempel's lucid and careful delineations of possibilities and constraints will remain pertinent.

Explanation

In the early decades of the twentieth century, the thought that one of the aims of the sciences is to provide explanations, traditionally popular, suffered a temporary eclipse. Thinking that appeals to the explanatory power of a theory reflected purely subjective judgments, scholars writing about science tended to concentrate on the criterion of predictive success (see, for one among many examples, Pearson). Hempel played the leading role in restoring the respectability of the concept of scientific explanation. From his earliest discussions of the topic, he insists on the objective character of scientific explanation (see 1965: 234 (originally written in 1942); also 1966: ch. 5). Taking up a theme already sounded by earlier empiricists (for example John Stuart Mill), and perhaps as old as Aristotle, he suggests that explaining a fact, state, or event consists in showing why that fact, event, or state could have been expected to occur, given the laws of nature. The key to explanation is nomic expectability.

Hempel proposed that an explanation is an argument whose conclusion is a statement describing the phenomenon to be explained (this statement is the *explanandum*) and whose premises (the *explanans*) include at least one law of nature. Although his early writings concentrated on cases in which the argument is deductive, he was explicit, from the beginning, that some explanations are non-deductive arguments. He also took considerable pains to point out that, as actually given, explanations may not take the ideal form he specified. So, for example, historians develop explanatory narratives that are far from complete arguments, and yet, Hempel contended, the explanatory force of their work derives from the possibility of recognizing general laws of nature, which in combination with the claims they advance would yield a compelling argument for the explanandum.

The *Deductive-Nomological* (D-N) model of explanation can be encapsulated in a schema: deductive explanatory arguments take the form

$$\begin{array}{l} C_1, \ldots, C_m \\ \underline{L_1, \ldots, L_n} \\ E \end{array}$$

where the C_i are statements reporting particular facts, the L_j are laws of nature, whose presence is essential to the validity of the argument, and E is the explanandum. (Hempel's model does not require that there be any *C*s, although there must be at least one *L*; his presentations sometimes identify E as a statement of particular fact, but he allows for explanations of this form whose conclusions are laws, including probabilistic laws.) To be a genuine explanation, the premises of an argument fitting the schema must all be true. If one or more of the premises is not true, then the argument counts as a *potential* explanation.

In the 1940s, Hempel hoped to articulate the D-N model more precisely, and he proposed a formal explication of the notion of law and of deductive explanation (Hempel and Oppenheim 1948; see Hempel 1965: ch. 10). Unfortunately, this attempt proved vulnerable to trivializing counterexamples, and, in any event, Goodman's explorations

of laws, counterfactuals, and induction, convinced Hempel that no formal account of scientific laws could be given. Thus, throughout the 1950s and 1960s, his work on scientific explanation focused on showing how his preferred approach to explanation illuminated aspects of the natural and social sciences and how it could be extended to include non-deductive arguments.

The latter task was complicated by an important disanalogy between deductive and inductive arguments. Adding extra premises to a deductively valid argument preserves validity, but the incorporation of new information into an argument that is inductively strong may not only undermine the argument but even support a contrary conclusion. So, to take one of Hempel's own examples, to claim that Jones is suffering from a streptococcal infection and that he is being treated with penicillin, together with the probabilistic law that 99 percent of those treated with penicillin recover from such infections confers high probability on the conclusion that Jones will recover, but if we now learn that this particular streptococcal infection is penicillin-resistant then we have strong reasons for thinking that Jones will not recover.

Hempel's model of *Inductive-Statistical* Explanation (I-S) proposed that I-S explanations are arguments with true premises of the form:

$$\begin{array}{l} Pr(B|A) = r \\ Ac \\ \hline\hline Bc \end{array} [r]$$

Here the double line indicates that the premises bestow on the conclusion the probability r, which is supposed to be close to one. To block the problem of the "ambiguity of statistical explanation," Hempel imposes the "requirement of maximal specificity." If s is the conjunction of the premises of the explanation, and if k is a statement logically equivalent to the set, K, of accepted sentences, then if $s\&k$ implies that c belongs to a subset A^* of A, then $s\&k$ must also imply that $Pr(B|A^*) = r^*$ where $r^* = r$ unless the conditional probability of B on A^* is simply a matter of probability theory (as, for example, when A^* is the null set). This intricate condition is intended to require that we always employ the most specific probabilistic information we have, and, as Hempel explicitly noted, it introduces an unwelcome relativization into the account of explanation, for, unlike D-N explanations, I-S arguments only qualify as explanations relative to a particular state, K, of our knowledge.

The *Covering Law Model of Explanation*, comprising the D-N and I-S models, was enormously influential, not only restoring the respectability of the concept of explanation but also sparking methodological discussions in the social sciences. The many-sided character of Hempel's lucid discussions, especially in the title essay of *Aspects of Scientific Explanation*, provides a model for philosophical exploration of an important metascientific concept. Nonetheless, for all the subtlety of his treatment, Hempel's account is no longer widely accepted among philosophers of science (although it continues to be adopted in other philosophical debates and in the methodological reflections of natural and social scientists).

Some of the difficulties stem from problems with which Hempel struggled. The intricate requirement of maximum specificity proved inadequate to salvage the notion of

I-S explanation, yielding the unwelcome consequence that *bona fide* inductive explanations turn out to be tacitly deductive (Coffa 1974). Another difficulty of the model of probabilistic explanation lies in the fact that we can apparently explain events that are unlikely to occur: even though it may be improbable that an atomic nucleus will undergo a particular sequence of decay, we can still, it seems, use quantum physics to explain the rare occurrences that do take that path. Even in the realm of purely deductive explanation, there are formidable challenges. As Hempel himself noted, a derivation of Boyle's Law from the conjunction of this law with Kepler's laws would satisfy the D-N schema, even though any such derivation is explanatorily worthless. How do we distinguish such arguments from the explanatory derivations of laws, for example the derivation of Kepler's laws within Newtonian gravitational theory?

Perhaps the most severe problems came from a cluster of examples that showed how familiar asymmetries that occur in the context of causal judgments also affect our assessments of explanatory power. Suppose that a flagpole casts a shadow of a particular length. Using the law of rectilinear propagation of light, together with facts about the height of the pole and the elevation of the sun, it is possible to derive the length of the shadow, a derivation that fits the D-N model. So far, so good, since that particular derivation seems genuinely explanatory. The trouble is that we can also work in the opposite direction. Given the propagation law, the elevation of the sun and the length of the shadow, we can derive the height of the pole, and this derivation fits the D-N schema equally well (Bromberger 1966). A natural response to examples like this is to declare that opaque objects *produce* (or cast) shadows and that shadows do not *produce* the associated objects, so that there is a causal asymmetry unrepresented in the Hempelian schema.

Hempel's scattered remarks about the connection between explanation and causation present a clear picture of his position. Influenced by Humean worries about the notion of causation, he holds that our understanding of causal relations is grounded in our ability to subsume phenomena under lawlike regularities. The concept of explanation is prior to that of causation, in that a claim that *c* caused *e* is always derivative from the thought that the occurrence of *e* would be properly explained by an argument in which a description of *c* figured among the premises, an argument satisfying the covering-law model. Hence, Hempel cannot appeal to causal asymmetries to reformulate his account of explanation, and, on the few occasions on which he confronts examples that embody such asymmetries, he argues strenuously that our intuitive responses to these cases should not be trusted (for example, 1965: 352–3). With the articulation of a family of instances like that of the shadow-casting flagpole, attempts to deny differences in explanatory worth came to appear ever more desperate, and most philosophers of science (including Hempel himself) have concluded that the problem of explanatory asymmetry cannot be dismissed as illusory.

For about a quarter of a century Hempel's account of scientific explanation almost achieved philosophical consensus. Since it succumbed to a host of problems and criticisms no successor approach has garnered similar support. Inspired by the problem of asymmetry, several philosophers have offered accounts that invoke the concept of causation (see, for example, Humphreys 1989, and Salmon 1984, 1998). Others have tried to preserve the main features of Hempel's account by developing an idea that received

passing attention in his own writings, and suggesting that explanation consists in the unification of the phenomena (Friedman 1974, Kitcher 1981). Yet others have contended that explanation is an activity whose crucial properties vary with context (Achinstein 1983, van Fraassen 1980: ch. 5). All the existing accounts face major obstacles (often gleefully noted by the partisans of rival accounts). If there is a consensus, its central tendency is that, while Hempel's covering-law model is inadequate, it is exemplary in demonstrating the range, rigor, and clarity that any satisfactory theory of explanation should strive for.

Hempel's legacies

In the past decades, logical empiricism has been criticized for various shortcomings: neglect of the historical development of science (Kuhn 1962/1970), overemphasis on the search for lawlike regularities in nature (Cartwright 1983, 1999), and failure to appreciate the autonomy of experimental practice (Galison 1987, Hacking 1983). At the same time, many philosophers have proposed that Hempel and his co-workers adopted an unnecessarily restrictive view of the formal resources on which accounts of confirmation, theories, and explanation might draw. Despite these complaints, philosophy of science continues to pursue the agenda that Hempel so lucidly articulated, and, if the set of questions has been enlarged and the Hempelian answers are no longer widely accepted, it would be foolhardy to tackle these problems without thorough awareness of Hempel's many insights.

No essay on Carl ("Peter") Hempel would be complete without some recognition of his extraordinary pedagogical influence. Not only was he the author of one of the great introductions to any field of philosophy (Hempel 1966), but, through lectures and seminars, he was an inspiration to generations of undergraduates, graduate students, and younger philosophers. Those who knew him saw, again and again, a rare combination of high scholarly integrity and personal kindness, acute intelligence and gentleness, and his daily actions reminded those around him that philosophy began with the desire for wisdom and for understanding the good. In his life, as well as in his work, Hempel was a true philosopher.

Bibliography

Works by Hempel

1936 (with Oppenheim, P.): *Der Typusbegriff im Lichte der neuen Logik*, Leiden: Sitjhoff.

1945a: "Geometry and Empirical Science," *American Mathematical Monthly* 52, pp. 7–17.

1945b: "On the Nature of Mathematical Truth," *American Mathematical Monthly* 52, pp. 543–56.

1948 (with Oppenheim, P.): "Studies in the Logic of Confirmation," *Philosophy of Science* 15, pp. 135–75. (Reprinted in Hempel 1965, ch. 10.)

1958: "The Theoretician's Dilemma," in *Minnesota Studies in the Philosophy of Science* II, ed. H. Feigl, M. Scriven, and G. Maxwell, Minneapolis: University of Minnesota Press, pp. 37–98. (Reprinted in Hempel 1965, ch. 8.)

1965: *Aspects of Scientific Explanation*, New York: Free Press.

1966: *Philosophy of Natural Science*, Englewood Cliffs, NJ: Prentice-Hall.

Works by other authors

Achinstein, P. (1983) *The Nature of Explanation*, New York: Oxford University Press.
Bromberger, S. (1966) "Why-Questions," in *Mind and Cosmos*, ed. R. Colodny, Pittsburgh: University of Pittsburgh Press.
Carnap, R. (1956) "The Methodological Character of Theoretical Concepts," in *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, ed. H. Feigl and M. Scriven, Minneapolis: University of Minnesota Press, pp. 38–76.
Cartwright, N. (1983) *How the Laws of Physics Lie*, Oxford: Oxford University Press.
——(1999) *The Dappled World*, Cambridge: Cambridge University Press.
Coffa, J. A. (1974) "Hempel's Ambiguity," *Synthese* 28, pp. 141–63.
Friedman, M. (1974) "Explanation and Scientific Understanding," *Journal of Philosophy* 71, pp. 5–19.
Galison, P. (1987) *How Experiments End*, Chicago: University of Chicago Press.
Glymour, C. (1980) *Theory and Evidence*, Princeton, NJ: Princeton University Press.
Goodman, N. (1949) "The Logical Simplicity of Predicates," *Journal of Symbolic Logic* 14.
——(1955) *Fact, Fiction, and Forecast*, Indianapolis: Bobbs-Merrill.
Hacking, I. (1983) *Representing and Intervening*, Cambridge: Cambridge University Press.
Humphreys, P. (1989) *The Chances of Explanation*, Princeton, NJ: Princeton University Press.
Kitcher, P. (1981) "Explanatory Unification," *Philosophy of Science* 48, pp. 507–31.
Kripke, S. (1971) "Naming and Necessity," in *Semantics of Natural Languages*, ed. D. Davidson and G. Harman, Dordrecht: Reidel.
Kuhn, T. S. (1970) *The Structure of Scientific Revolutions,* Chicago: University of Chicago Press. (First published 1962.)
Nagel, E. (1962) *The Structure of Science,* New York: Harcourt Brace.
Pearson, K. (1911) *The Grammar of Science,* 3rd edn., London: A. & C. Black.
Putnam, H. (1973) "Meaning and Reference," *Journal of Philosophy* 70, pp. 699–711.
Quine, W. V. (1953) *From a Logical Point of View*, Cambridge, MA: Harvard University Press.
Reichenbach, H. (1938) *Experience and Prediction*, Chicago: University of Chicago Press.
Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World,* Princeton, NJ: Princeton University Press.
——(1998) *Causality and Explanation,* New York: Oxford University Press.
Suppe, F. (ed.) (1970) *The Structure of Scientific Theories,* Urbana, IL: University of Illinois Press.
Suppes, P. (1967) "What is a Scientific Theory?," in *Philosophy of Science Today*, ed. A. Danto and S. Morgenbesser, New York: Basic Books.
van Fraassen, B. (1980) *The Scientific Image*, Oxford: Oxford University Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

12

Nelson Goodman (1906–1998)

ISRAEL SCHEFFLER

Nelson Goodman, distinguished American philosopher, was born on August 7, 1906 in Somerville, Massachusetts, and died in Needham, Massachusetts on November 25, 1998. He received the Bachelor of Science degree from Harvard University in 1928 and the Ph.D. from Harvard in 1941. From 1929 to 1940 he operated an art gallery in Boston; throughout his life, he was a collector of ancient and modern art. From 1942 to 1945 he served in the United States Army. Thereafter, he taught for one year at Tufts University before his appointment to the faculty of the University of Pennsylvania, where he served as associate professor from 1946 to 1951, and then as professor from 1951 to 1964. From 1964 to 1967 he was the Harry Austryn Wolfson Professor of Philosophy at Brandeis University. From 1968 to 1977, he was professor of philosophy at Harvard University.

Goodman's contributions to philosophy are wide-ranging, penetrating, and fundamental. The areas in which he worked include epistemology, philosophy of science, philosophy of language, analysis of simplicity, theory of symbols, aesthetics, and metaphysics. His work is characterized by unusual originality, typically rejecting conventional approaches in order to reconceive the problems to be addressed and then proposing provocative solutions to them. Thus, for example, he recasts the traditional problem of induction so as to require codification, rather than justification of inductive practice, thereafter dooming the prospects of purely syntactic or semantic approaches to such codification and offering a new pragmatic treatment based on "entrenchment." To take another example, he reorients aesthetics as a division of epistemology, concerned primarily with understanding rather than evaluating works of art, a project that leads him to formulate a comprehensive new theory of referential functions embracing the literary, pictorial, and other arts as well as the sciences. Finally, rejecting both physicalism and phenomenalism, both realism and idealism, he emphasizes the diversity of equally adequate conflicting conceptualizations for any subject matter, thus championing what he calls "irrealism"; the doctrine that there are many worlds if any and that worlds are made, not found.

Goodman's treatments are analytically subtle as well as inventive. His writing is terse and telling, making brilliant use both of logic and of metaphor. To summarize his work in this brief space is clearly impossible. Instead, we shall give a selective account of his main contributions in basic areas of his thought.

Likeness of meaning

Goodman's discussions with W. V. Quine and Morton White in the late 1940s led to a widespread discrediting of the analytic/synthetic distinction. Goodman's paper, "On Likeness of Meaning" reflects some of his contribution to these discussions. "Under what conditions," he asks, "do two names or predicates in an ordinary language have the same meaning?" He considers and rejects such answers to the question as that they stand for the same essence or the same image or idea, or that nothing can be conceived that satisfies the one but not the other, or that it is impossible that something satisfies the one but not the other. Eschewing all reference to essences, images, ideas, conceivability, and possibility, he asks whether two predicates have the same meaning if and only if they are coextensive, that is, apply to exactly the same things. The answer is no, since there are clear cases where words with the same extension (or equally lacking in it) differ in meaning, for example "centaur" and "unicorn." Extensional identity is indeed a necessary, but not a sufficient condition for sameness of meaning.

Here, Goodman proposes that it is not only the extensions of the original two words themselves that we need to consider (so-called *primary* extensions) but also the extensions of their parallel compounds (so-called *secondary* extensions). A pair of parallel compounds is formed by making an identical addition to each of the two words under consideration; thus, adding the word "picture" to "centaur" and to "unicorn," we have the parallel pair "centaur-picture" and "unicorn-picture." Now, although there are neither centaurs nor unicorns, there certainly are centaur-pictures and unicorn-pictures and, moreover, they are different. Although the original words have the same extension, the parallel compounds differ in extension. Goodman's idea is, then, that the difference in meaning between two words is a matter either of their own difference in extension or of that of any of their parallel compounds. In general, terms have the same meaning if and only if they have the same primary and secondary extensions.

This idea is generalized to cover cases in which the addition of "picture" yields a compound with null extension. For example, "acrid-odor-picture" and "pungent-odor-picture" have the same (null) extension, neither applying to anything. Compounds can, however, be formed by other additions, and Goodman suggests that "description" constitutes a suffix capable of yielding all the wanted distinctions for every pair of words *P* and *Q*. For any inscription of the form "a *P* that is not a *Q*" is a thing denoted by the compound "*P*-description" but not by the parallel "*Q*-description." And any inscription of "a *Q* that is not a *P*" belongs to the extension of "*Q*-description" but not to that of "*P*-description." Thus, "pungent-odor-description" and "acrid-odor-description" differ extensionally since the first, but not the second, applies to any inscription of the form "a pungent odor that is not an acrid odor," and vice versa. Thus, even if all pungent odors are acrid and acrid odors pungent, the terms "pungent odor" and "acrid odor" differ in meaning. It follows from this proposal, in fact, that no two different words have the same meaning.

To the objection that the compounds used in deriving this radical conclusion are "trivial," Goodman replies that when a single form of compound indeed has a different extension for every term, "the fact that it has different extensions for two given terms is of no striking or special interest. Let us, then, simply exclude every compound for which the corresponding compounds of every two terms have differing extensions . . .

instead of saying that every two terms differ in meaning but that some may not differ in interesting ways, we say that two terms differ in meaning only if they differ in certain interesting or peculiar ways" (1972: 237–8).

The new riddle of induction

The starting point for modern discussions of induction is Hume's denial of necessary connections of matters of fact. Effects cannot be simply deduced from their causes, nor can predictions be logically demonstrated on the basis of available evidence garnered from past experience. What then can be the rational justification of the predictions upon which we base all our actions? Hume answers that while there is indeed no deductive justification, there is a mental habit which underlies the expectation that phenomena uniformly conjoined in our past experience will be so conjoined in the future. In effect, he offers the uniform past conjunction of events as a mark of those inductions we find compelling in making our predictions. This idea is also represented by a modern version, which has found wide favor among scientists as well as philosophers. According to this version, predictions are made in conformity with generalizations that have regularly worked in the past. Such congruence with past experience is of course no guarantee of future success, but it seems to single out those predictions we adopt at any given time. Lacking such guarantee, however, what justification can there be for adopting these predictions?

Goodman argues that the justification of *induction*, like that of *deduction*, is only a matter of codifying our particular sanctioned inferences, and coordinating them with the governing rules of our practice, thus bringing them into agreement with one another. "A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either" (1983: 64).

How persuasive, then, in codifying our practice is the prevalent regularity doctrine, the view that our sanctioned predictions are those in conformity with generalizations that have regularly worked so far? Here Goodman introduces the notorious case of the green emeralds: suppose all emeralds examined before the present time t are green. We predict the next emerald to be examined will be green, since such prediction conforms to the generalization that all emeralds are green, a generalization uniformly confirmed by all our past evidence. Consider now the predicate "grue," applicable to everything examined before t if and only if green but to everything else if and only if blue. Then, all emeralds examined before t are not only green, but also grue. Hence, the generalization that all emeralds whatever are grue is supported by no less evidence than the generalization that they are all green. The prediction that the next emerald to be examined will be blue is thus, by the regularity theory, as confirmed as the prediction that it will be green. This theory, whether in Hume or in modern scientific dress, thus fails utterly to separate the properly confirmed "green" prediction from the bogus "grue" one. "Regularity in greenness confirms the prediction of further cases; regularity in grueness does not. To say that valid predictions are those based on past regularities, without being able to say *which* regularities, is thus quite pointless. Regularities are

where you find them, and you can find them anywhere" (1983: 82). If the old problem of justifying our inductive practice has indeed been supplanted by "the new riddle of induction" asking for a principled distinction between valid and invalid predictions (or more generally, "projections"), we still have a long way to go.

Goodman's solution depends on utilizing knowledge typically not used in attempts to interpret induction. In particular, he presumes some knowledge of past projections, that is, of hypotheses that have been actually projected in the past, "adopted after some of [their] instances have been examined and determined to be true, and before the rest have been examined" (1983: 87). Now, when we consult the record of past projections, we find that "green" has clearly been projected much more often than "grue"; it is much better *entrenched* than the latter. The entrenchment of a predicate flows from the actual past projections of it and of all other coextensive predicates. While in a sense it is thus the class that is entrenched, it becomes so only through the projection of terms that determine it.

Goodman, upon this basis, elaborates a subtle general theory of projection which would take us too far afield to characterize here. But the main point to be noted is that he hopes to have gone beyond the regularity theory by appealing to regularities in our linguistic habits. "Like Hume, we are appealing here to past recurrences, but to recurrences in the explicit use of terms as well as to recurrent features of what is observed. Somewhat like Kant, we are saying that inductive validity depends not only upon what is presented but also upon how it is organized; but the organization we point to is effected by the use of language and is not attributed to anything inevitable or immutable in the nature of human cognition. To speak very loosely, I might say that in answer to the question what distinguishes those recurrent features of experience that underlie valid projections from those that do not, I am suggesting that the former are those features for which we have adopted predicates that we have habitually projected" (1983: 97).

Constructionalism and nominalism

In his approach to philosophy, Goodman was greatly influenced by the constructionalism exhibited in Rudolf Carnap's *Der logische Aufbau der Welt*, whereby Carnap strove to reduce all concepts to the immediately given, with the help of the new mathematical logic (see CARNAP). Indeed, in his monumental *The Structure of Appearance*, Goodman includes a detailed exposition and critique of Carnap's system before developing an alternative phenomenalistic system taking qualia as atoms rather than Carnap's *elementarerlebnisse*.

In addition to offering acute analyses of qualia and their concretion, of size, shape, order, measure, and time, Goodman devotes the first part of his book to the theory of constructional systems. Here he argues that not only is intensional identity too strong as a requirement for constructional definition; but even extensional identity is too strong, since it precludes alternative adequate systematizations, e.g. taking points as certain classes of volumes or taking points as certain pairs of lines. He proposes instead a criterion of "extensional isomorphism," which he explains as allowing a given term to be definable alternatively "by any of several others that are not extensionally identical with one another" (1977: 17). The criterion, which must be applied to the whole

set of definitions of the system, nevertheless provides for the truth-value preserving character of the translations of all sentences we care about.

In his constructionalism, Goodman is concerned both with system and with simplicity. Indeed, these two concerns are, as he argues, identical. "The purpose of constructing a system is to interrelate its predicates. The same purpose is served by reducing to a minimum the basis required. Every definition at once both increases the coherence of the system and diminishes the number of predicates that need to be taken as primitive. Thus the motive for seeking economy is not mere concern for superficial neatness. To economize and to systematize are the same" (1977: 48).

Goodman's attitude toward constructional systems issues in a passionate defense of the *Aufbau* against the charge that it is abstract, static, and bloodless, a mere caricature of experience. On the contrary, such a charge is in effect an attack against philosophy in general, for all philosophy involves conceptualization, abstraction, and systematization, the effort to map experience, not to duplicate it. "A map is schematic, selective, conventional, condensed, and uniform . . . The map not only summarizes, clarifies, and systematizes, it often discloses facts we could hardly learn immediately from our explorations." Goodman emphasizes the fact that different maps are useful for different purposes. "Let no one suppose that if a map made according to one scheme of projection is accurate then maps made according to alternative schemes are wrong" (1972: 15–16). This point is worth remarking since Goodman's phenomenalistic basis in *The Structure of Appearance* is not, as has often been mistakenly supposed, a matter of philosophical conviction precluding other, for example, physicalistic bases; he insists that while a constructional system may be adequate and illuminating, it can by no means claim a monopoly of wisdom. And he sees no virtue in claims of epistemological priority made on behalf of either phenomenalistic or physicalistic systems.

In constructing systems Goodman insists that the logic employed is not mere neutral machinery. Interpreted use of the calculus of classes – employing variables taking classes as values – commits the system not only to the individuals explicitly acknowledged, but also to classes of classes, etc. of these, without limit. "The nominalistically minded philosopher like myself," he declares, "will not willingly use apparatus that peoples his world with a host of ethereal, platonic, pseudo-entities. As a result, he will so far as he can avoid all use of the calculus of classes, and every other reference to non-individuals, in constructing a system" (1977: 26). Goodman's principle is "entities differ only if their content at least partially differs." He offers the following example: "A class (e.g. that of the counties of Utah) is different neither from the single individual (the whole state of Utah) that exactly contains its members nor from any other class (e.g. that of acres of Utah) whose members exactly exhaust this same whole" (1977: 26).

A pioneering paper by Goodman and W. V. Quine is "Steps Toward a Constructive Nominalism" (1947) in which the authors, starting from a renunciation of abstract entities, proceed to offer ingenious nominalistic translations of a variety of non-nominalistic statements (see QUINE). They soon, however, concede their inability thus to translate all of mathematics and they therefore suggest a different way of saving mathematics nominalistically: to devise a nominalistic syntax capable of describing the syntactic rules by which mathematical inscriptions, as concrete uninterpreted marks, are manipulated.

In this enterprise, they succeed in defining a suitable nominalistic syntax language for describing the object language of mathematics, with its ingredient notions of "axiom," "rule," "proof," and "theorem." In conclusion, they declare "Our position is that the formulas of platonistic mathematics are, like the beads of an abacus, convenient computational aids which need involve no question of truth. What is meaningful and true in the case of platonistic mathematics as in the case of the abacus is not the apparatus itself, but only the description of it: the rules by which it is constructed and run. These rules we do understand, in the strict sense that we can express them in purely nominalistic language" (1972: 198).

Further discussion of nominalism, including the later differing interpretations of Goodman and Quine, as well as elaborations and answers to various objections, can be found in Goodman's paper, "A World of Individuals" (1972: 155–72). Here, he declares that his nominalism is specifically "the refusal to recognize classes" (p. 156), requiring that all entities admitted to a system, no matter what they are, be treated as individuals.

Theory of symbols

In *Languages of Art*, Goodman develops a theory of symbols that encompasses but ranges far beyond what is typically considered aesthetics, while giving only minor attention to artistic value and criticism. His primary concern is to develop a systematic approach to the functioning of symbols not only in the arts but also in the sciences and in ordinary contexts as well, for example, in the everyday use of labels and samples. And in the realm of the arts proper, he discusses music as well as painting, literary arts as well as dance and architecture. His key emphasis is on cognitive function and to this end he devises a general theory which he employs in characterizing notations, sketches, scripts, and paintings. In related discussions, his book offers an account of forgery of works of art, a theory of metaphor and a categorization of figures of speech, a treatment of fictional expressions, and a comprehensive interpretation of reference as including not only denotation, but also exemplification and expression.

A swatch of cloth in a tailor's shop is typically used to exemplify certain of its properties, i.e. its color and texture, but not its size or shape. To serve thus, it must possess these properties, but also refer to them. *Exemplification* is a symbolic function by which the sample stands for a property it possesses. *Expression* implies metaphorical exemplification: if a picture expresses sadness, it is itself metaphorically sad and also refers to sadness. With these two notions at hand, Goodman is able to treat works of art not merely as objects of reference but – even where nonrepresentational – as referring symbols in their own right. And the ingredient idea of metaphor vastly expands the resources available for treating the wide expressive ranges of works of art.

Goodman's treatment of metaphor has been widely noted. "Where there is metaphor," he writes, "there is conflict: the picture is sad rather than gay even though it is insentient and hence neither sad nor gay. Application of a term is metaphorical only if to some extent contra-indicated" (1976: 69). Along with such contra-indication, there must also be attraction, or aptness. The metaphorical use of "sad" implies that there are two ranges for the term "sad," but that these two ranges do not simply comprise an ambiguity. In mere ambiguity, the separate uses of the term are

independent. "In metaphor, on the other hand, a term with an extension established by habit is applied elsewhere under the influence of that habit; there is both departure from and deference to precedent. When one use of a term precedes and informs another, the second is the metaphorical one" (p. 71).

The key technical discussion in the book presents a theory of notation, for example, a musical score. The main function of such a notation, says Goodman, is to provide authoritative identification of a work from performance to performance, whatever other uses it may have. "What is required is that all and only performances that comply with the score be performances of the work" (1976: 128). With this principle as a guide, Goodman develops five requirements for a notational system: unambiguity and syntactic and semantic disjointness and differentiation.

To give a precise account of these requirements would exceed the limits of this space, but a brief summation of their import is stated by Goodman as follows: "A system is notational, then, if and only if all objects complying with inscriptions of a given character belong to the same compliance class and we can, theoretically, determine that each mark belongs to, and each object complies with inscriptions of, at most one particular character" (1976: 156). Using this notion of notation, Goodman is able then to characterize other things than, for example, scores. A script, for example "is a character in a notational *scheme* and in a language but, unlike a score, is not in a notational *system*. The syntactic but not all the semantic requirements are met . . . [it is] a character in a language that is either ambiguous or lacks semantic disjointness or differentiation" (pp. 199, 201).

Rather than offering a definition of the aesthetic, Goodman suggests four symptoms that tend to be present in aesthetic experience: syntactic density, semantic density, syntactic repleteness, and exemplificational character. The first two call for maximum sensitivity of discrimination, the third calls for such effort along many dimensions, while the fourth is shown by concern with properties exemplified by a symbol, not merely with things the symbol denotes. Largely as a result of his analysis, Goodman emphasizes the affinities between art and science, despite their differences. The difference between them is not that between feeling and fact or truth and beauty "but rather a difference in domination of certain specific characteristics of symbols" (1976: 264).

Irrealism

In Goodman's *Ways of Worldmaking*, he speaks of versions as including depictions as well as descriptions, works of art as well as works of science. And he insists that there are conflicting right world-versions, rather than a single world underlying the rightness of all. He considers his view on these matters as

> belonging in that mainstream of modern philosophy that began when Kant exchanged the structure of the world for the structure of the mind, continued when C. I. Lewis exchanged the structure of the mind for the structure of concepts, and that now proceeds to exchange the structure of concepts for the structure of the several symbol systems of the sciences, philosophy, the arts, perception, and everyday discourse. The movement is from unique truth and a world fixed and found to a diversity of right and even conflicting versions or worlds in the making.

The view that emerges, he says, can perhaps be described as "a radical relativism under rigorous restraints, that eventuates in something akin to irrealism" (1978: x).

Goodman takes his adversary to be "the monopolistic materialist or physicalist who maintains that one system, physics, is preeminent and all-inclusive, such that every other version must eventually be reduced to it or rejected as false or meaningless" (1978: 4). Were all versions reducible to a single one, that one might be plausibly considered the only truth about the one world. But such reducibility is a chimera; the claim of reducibility to physics is "nebulous since physics itself is fragmentary and unstable and the kind and consequences of reduction envisaged are vague . . . The pluralist's acceptance of versions other than physics implies no relaxation of rigor but a recognition that standards different from yet no less exacting than those applied in science are appropriate for appraising what is conveyed in perceptual or pictorial or literary versions" (p. 5).

Doesn't the rightness of these various versions imply an underlying world that makes versions right? No, says Goodman,

> we might better say that "the world" depends on rightness. We cannot test a version by comparing it with a world undescribed . . . all we learn about the world is contained in right versions of it; and while the underlying world, bereft of these, need not be denied to those who love it, it is perhaps on the whole a world well lost . . . For many purposes . . . just versions can be treated as our worlds. (1978: 4)

Worlds are made through the making of versions, but versions cannot be made any way we like. Goodman denies that his acceptance of many right world-versions implies "that anything goes, that tall stories are as good as short ones, that truths are no longer distinguished from falsehoods." Although it is true, he says, that "we make worlds by making versions," we cannot do so at random or by whim (1978: 94). "Of course," he says,

> we want to distinguish between versions that do and those that do not refer, and to talk about the things and worlds, if any, referred to; but these things and worlds and even the stuff they are made of – matter, anti-matter, mind, energy, or whatnot – are fashioned along with the things and worlds themselves. (p. 94)

The making of worlds is brought about by the making of versions and the multiple worlds thus made "are just the actual worlds . . . answering to true or right versions. Worlds possible or impossible supposedly answering to false versions," says Goodman, "have no place in my philosophy" (1978: 94).

Goodman's relativism on the issue of worldmaking is familiar from his earlier work on constructionalism, where, as we have seen, he insists on honoring the conflicting definitions that can be offered for the very same terms. The criterion of extensional isomorphism, as we have noted above, is a more relaxed standard than either synonymy or identity, but it nonetheless imposes clear and definite restraints.

Similarly, his theory of worldmaking, paradoxical as it may seem, recognizes a vast array of conflicting versions, rather than a single underlying world, yet insists that there are clear distinctions to be drawn between right and wrong versions. Goodman's

relativism must therefore be sharply distinguished from nihilism, subjectivism, and cultural relativism. His hospitality to variant and opposed conceptualizations is allied with a dedication to the highest standards of logical rigor.

Bibliography

Works by Goodman

1947 (with Quine, W. V.): "Steps Toward a Constructive Nominalism," *Journal of Symbolic Logic* 12, pp. 105–22. (Reprinted in Goodman 1972, pp. 173–93.)
1949: "On Likeness of Meaning," *Analysis* 10, pp. 1–7. (Reprinted in Goodman 1972, pp. 221–30.)
1953: "On Some Differences about Meaning," *Analysis* 13, pp. 90–6. (Reprinted in Goodman 1972, pp. 231–8.)
1963: "The Significance of *Der logische Aufbau der Welt*," in *The Philosophy of Rudolf Carnap*, ed. P. A. Schilpp, La Salle, IL: Open Court and London: Cambridge University Press. (Reprinted in Goodman 1972, pp. 5–23.)
1972: *Problems and Projects*, Indianapolis: Bobbs-Merrill.
1976: *Languages of Art*, 2nd edn., Indianapolis: Hackett. (First published Indianapolis: Bobbs-Merrill, 1968.)
1977: *The Structure of Appearance*, 3rd edn., Dordrecht: Reidel. (First published Cambridge, MA: Harvard University Press, 1951.)
1978: *Ways of Worldmaking*, Indianapolis: Hackett.
1983: *Fact, Fiction and Forecast*, 4th edn., Cambridge, MA: Harvard University Press. (First published 1955 by Harvard University Press.)
1984: *Of Mind and Other Matters*, Cambridge, MA: Harvard University Press.
1988 (with Elgin, Catherine Z.): *Reconceptions in Philosophy and Other Arts and Sciences*, Indianapolis: Hackett.

Works by other authors

Elgin, C. Z. (1983) *With Reference to Reference*, Indianapolis: Hackett.
——(ed.) (1997) *The Philosophy of Nelson Goodman*, 4 vols, New York and London: Garland Publishing. (Contains articles by various authors on Goodman's constructionalism, theory of induction, philosophy of art, and theory of symbols.)
Howard, V. A. (1982) *Artistry*, Indianapolis: Hackett.
McCormick, P. J. (ed.) (1996) *Starmaking*, Cambridge, MA: MIT Press. (Contains critical discussions of Goodman's irrealism, by Hempel, Putnam, Scheffler, and Goodman.)
Scheffler, I. (1963) *The Anatomy of Inquiry*, New York: Alfred A. Knopf.
——(1997) *Symbolic Worlds, Art, Science, Language, Ritual*, Cambridge: Cambridge University Press.
Stalker, D. (1994) *Grue: The New Riddle of Induction*, Chicago: Open Court.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

13

H. L. A. Hart (1907–1992)

SCOTT SHAPIRO

Herbert Lionel Adolphus Hart was born of Jewish parents in Yorkshire, England and was educated at New College, Oxford. After graduating with a First in Greats, Hart was called to the Bar as a Chancery barrister in London. He spent the next eight years building a successful legal practice, specializing first in property conveyancing, trust drafting, and tax planning, and then moving on to court work and advising. Although his interests quickly turned from law to philosophy, Hart continued to practice and, in fact, during this period he declined an invitation to teach philosophy in Oxford. His legal career, however, was cut short by World War II. While working with British Intelligence, Hart met Gilbert Ryle and Stuart Hampshire, from whom he learned of the new trends in philosophy. When the war ended, Hart left his law practice and returned to Oxford. In 1952, Hart was elected to the Chair of Jurisprudence, a somewhat surprising appointment given that he did not have a degree in either law or philosophy and had published little by that point. He occupied that chair, however, with great distinction, publishing several seminal works in legal theory, including his masterpiece, *The Concept of Law*, in 1961.

Hart is perhaps best known for his vigorous and sophisticated defense of the doctrine known as legal positivism. In its broadest sense, legal positivism is a theory about the nature of law that denies any necessary connection between legality and morality. No stipulation is made that, in order to count as law, a norm must possess any moral attributes. Legal positivists, therefore, believe that it is possible for a legal system to recognize a rule as legally valid even if it happens to be unjust. This analytic separation between the legal and the moral was captured by John Austin when he said: "The existence of law is one thing; its merit or demerit is another" (1954: 184–5).

In an effort to cleanse analytic jurisprudence of its moral content, every legal positivist before Hart thought it necessary to recast the basic legal concepts of *obligation*, *rule*, *validity*, and *authority* in terms of sanctions. Austin, for example, believed that legal rules are nothing more than orders backed by threats of sanctions issued by the sovereign. Sovereignty, in turn, was understood in terms of coercive power, the sovereign being the one in a group who has the power to elicit habitual obedience from every one and who habitually obeys no one. Someone is under a legal obligation to act, on this view, if they are likely to be sanctioned for failing to act.

Hart was firmly committed to the analytic separation of law and morality, but thought that these sanction-centered theories distorted and concealed the various ways in which the law guides conduct. For example, Hart pointed out that there are many legal rules that lack sanctions, in the sense that no penalties are imposed as a result of non-conformity with them. If a person drafts a will but fails to have it signed by two witnesses, that person is not sanctioned for the inadequate attention paid to the testamentary rules. He has simply failed to form a valid will and his actions lack legal effect.

Sanction-centered theories fall short, according to Hart, because they treat all legal rules as if their sole function is to discourage undesirable behavior. Their paradigm is the criminal law, where the rules impose duties to act or forbear from certain behavior and specify sanctions in the event of disobedience. However, not only do the rules related to valid will or contract formation lack sanctions, but, as Hart observed, it does not even make sense to speak of obeying or disobeying them. These rules do not impose *duties*; they instead confer *powers*. Their function is not to discourage people from acting in ways that they otherwise might wish, but to give them facilities for realizing their wishes.

The effacement of power-conferring rules is especially problematic with respect to those rules that confer legal powers on public officials. Without such rules, Hart noted, sanction-centered theories cannot account for the *self-regulating* nature of legal institutions: it is a defining feature of law, as opposed to pre-legal social systems, that its officials are empowered to change the rules and to resolve the disputes that may arise under them.

Hart also believed that these theories give a misleading picture of the nature of the law's normativity. On the sanction-centered approach, the only reasons for action that the law provides are threats of sanctions. This ignores what Hart called *the internal point of view*, which is the perspective of those who treat the rules as standards of acceptable conduct. In every legal system, Hart claimed, some members of the group treat the rules not just as threats, or predictions of what courts will do, but as guides to their conduct and standards for the evaluation of others – as norms that *obligate* and *empower*, not merely *oblige.*

By emphasizing the internal point of view, Hart was not simply criticizing fellow legal positivists for neglecting an obvious fact, i.e. that at least some people in some circumstances are motivated by the law *qua* law, instead of sanctions. Rather, Hart was also mounting a methodological offensive against the crude scientism of some of his contemporaries. For example, Alf Ross, the Scandinavian Legal Realist, based his legal positivism on his commitment to logical positivism and believed that, if jurisprudence is to have empirical content, legal concepts must be operationalized in purely behavioristic terms. By contrast, Hart believed that theories of law must make essential reference to the attitudes of legal actors because the law is a *social practice*. In order to analyze the practice, it is not enough to record regularities of behavior; one must understand how the participants understand it. Hart's introduction of the internal point of view thus inaugurated the *hermeneutic turn* in jurisprudence, where the law is studied from the *inside*, that is, from the perspective of those who live under, and directly experience, the law.

By engaging in this hermeneutic enterprise, Hart was not, however, giving up on a naturalistic approach to legal theory. Indeed, Hart believed that the internal point of

view allowed the legal positivist to anchor rules in social facts. According to Hart, a social rule in a community exists whenever a sufficient number of people engage in a practice from the internal point of view. This internal aspect of rules is manifested externally in conforming behavior, as well as criticisms that attend deviations from the practice and the use of normative language such as *ought*, *must*, and *obligation* to express such disapprobation. The existence of a rule, therefore, is firmly rooted in the natural world, that is, in regularities of behavior motivated by the appropriate critical attitude.

Hart's theory of social rules forms the foundation of his approach to law. According to Hart, at the root of every legal system is a social rule of a special sort, which he called *the rule of recognition.* This rule imposes a duty on courts to apply rules that bear certain characteristics. In the American system, for example, the rule of recognition requires judges to apply the rules duly enacted by Congress. The rule of recognition, therefore, sets out the criteria of legal validity, that is, those criteria that a rule must possess in order to be law.

The rule of recognition is what Hart called a *secondary* rule: it is a rule about other rules. It is also an *ultimate* rule: it exists because it is accepted by judges from the internal point of view, not in virtue of its validation by another rule. The primary rules, by contrast, owe their existence to the rule of recognition, and not to any guidance that they might engender.

In addition to the rule of recognition, Hart argued that every legal system contains two other secondary rules. The *rule of change* confers the power on legislative bodies to modify the primary rules, whereas the *rule of adjudication* confers the power on courts to adjudicate whether the primary rules have been followed or violated.

By understanding the law as the union of primary and secondary rules, Hart introduced what might be called a *rule-centered* theory of the law. On this model, the law guides conduct not by issuing naked threats, but by providing rules that impose duties and confer powers. The basic legal concepts are also understood in terms of rules, not sanctions. A rule is valid in a legal system when the rule bears those characteristics set out in that system's rule of recognition. An act is legally obligatory, in turn, when it is required by a legally valid rule. A person has supreme legal authority when the secondary rules of the system confer legal power on that person and no other has been conferred a greater power. Even the concept of a sanction is rendered in terms of rules, for a sanction is not simply a cost imposed by the law, but, unlike a tax, is a penalty exacted because a rule has been violated.

Hart did not think that privileging the concept of a rule compromised the analytic separation of law and morals. In his model, a primary legal rule exists just in case it is validated by that system's rule of recognition. There is no demand that the criteria of legal validity set out make reference to the rule's moral properties. It is possible, and regrettably often the case, that a legal rule exists even though, from a moral point of view, it should not. And while it is true that the rule of recognition would not exist unless judges accept it from the internal point of view, this does not mean that they judge it *morally* acceptable or that it is morally acceptable for them to treat it in this way.

Despite Hart's insistence that law be seen as a system of rules, he did not think that judges are always guided by these rules when they decide cases. In his view, courts are not simply the passive servants of the legislature or of tradition, restricted to applying

the rules laid down in advance, but are active players in the creation and development of the law. Judges do not always *find* the law; they sometimes *make* it as well.

Hart, however, was not disturbed by the fact of judicial legislation. He thought that judges should be given free rein to decide some cases, as it enables them to fashion sensible solutions to unforeseen problems. Moreover, given the inherent limitation of natural languages, he believed that judicial legislation was unavoidable. According to Hart, all general terms in natural language (e.g. *vehicle*) contain a core of settled meaning (e.g. *car*) and a penumbra where the reference class is ill-defined (e.g. *tractor*). When a case falls into the core of a general term of the rule, the rule applies and the judge is legally obligated to apply the rule. However, when in the penumbra, the law *runs out* and the judge must exercise his discretion. By necessity, the judge cannot find the law, because there is no law to find, and hence must make new law.

Although sounding sensible enough, Hart's recognition, and sympathetic acceptance, of judicial legislation has been attacked by his chief critic, and successor to the Chair in Jurisprudence, Ronald Dworkin. In Dworkin's view, the role of a judge is to vindicate the legal rights of the parties and this can only be accomplished if the law completely regulates the judge's behavior in every case. Dworkin faulted Hart for counting as law only rules that have social pedigrees, such as legislation or custom, and ignoring the mass of implicit law represented by moral principles that justify the pedigreed rules and that determine the legally correct answer when these rules run out.

By arguing that, in every case, there is a right answer, Dworkin was not only challenging Hart's theory of adjudication but also his claim that law and morality were conceptually distinct. For if the legally correct answer is determined in part by norms whose only claim to legal validity is their moral validity, then it would seem that morality would be a determinate of legality, contrary to legal positivistic strictures.

In the Postscript to the second edition of *The Concept of Law*, published posthumously, Hart agreed with Dworkin that judges are often legally obligated to apply moral principles that lack pedigrees, and that when judges act on them, they are applying existing law. However, Hart believed that such a position was consistent with legal positivism, for he saw no reason why the rule of recognition could not validate a norm based on its moral properties. Legal positivists, according to Hart, only claim that a rule of recognition *need not* validate a norm on the basis of its moral content, not that it *cannot*. Even when the rule of recognition did validate principles on the basis of their moral content, Hart doubted that these principles would indicate a unique result in every case, thus leaving ample room for the exercise of judicial discretion.

In separating law from morals, Hart did not mean to preclude moral criticism of the law. Quite the contrary, Hart was a vocal and influential critic of many aspects of the criminal law, especially the prohibitions on so-called *private vices.* In *Law, Liberty and Morality*, Hart attacked the doctrine known as *legal moralism*, the belief that society has the right to use the criminal law to enforce its moral code. Lord Devlin had argued that social cohesion is possible only when a common code of morality is respected by all, and the flouting of that code, even in private, threatens such cohesion and, in turn, society's very existence. Hart noted that Devlin failed to produce any evidence supporting his causal claims, and doubted whether any could be mustered. More importantly, he argued that a society that criminalizes behavior that the majority finds

offensive is not a society that respects liberty. To respect liberty, a society must protect the right of individuals to choose their own lifestyle, even when it does not approve of the lifestyle they end up choosing. The *liberty* to act only in ways that others like is, as Hart pointed out, liberty in name only.

In contrast to most of his contemporaries, Hart eschewed grand moral theories in favor of a more commonsense approach to normative analysis, which borrowed elements from both the Utilitarian and Kantian traditions. For example, Hart thought that the justifying aim of punishment is the deterrence of crime. Yet, he also believed that this pursuit must yield to the demands of justice, so that it is wrong to punish people for crimes they did not commit or could not have helped committing. He was thus critical of the attempts to increase the efficiency of the criminal law by eliminating many of the traditionally recognized excuses, such as the restrictions on the use of the insanity defense and the introduction of crimes of strict liability and negligence.

Although Hart recognized that the availability of excuses might allow some to feign incapacity or mistake and thus evade responsibility, he nevertheless thought that the costs are slight compared to the benefits. Not only is it fundamentally unfair to punish those who could not have helped doing what they did, but, as Hart pointed out, a system of excuses places individuals in control of their destinies. For when the law only punishes people for actions they can avoid, people can avoid being punished. As long as individuals never *choose* to break the rules, the law will let them go about their lives. As a result, individuals need not fear that they will unwittingly bring the wrath of the law down upon themselves; they can rely on the fact that the law will excuse behavior that was not, in some suitable sense, a product of choice.

It is a mistake, Hart concluded, to think that reducing crime by eliminating excuses will lead to an increase in security. When excuses are unacceptable, people are unable to predict the consequences of their actions. A world that is unknowable and uncontrollable is a world in which no one is secure.

Bibliography

Works by Hart

1955: "Are there Any Natural Rights?," *Philosophical Review* 64, pp. 175–91. (Reprinted in *Political Philosophy*, ed. A. Quinton, Oxford: Oxford University Press, 1967, pp. 53–66.)
1958: "Positivism and the Separation of Law and Morals," *Harvard Law Review* 71, pp. 593–629.
1959 (with Honore, A. M.): *Causation in the Law*, Oxford: Clarendon Press.
1961: *The Concept of Law*, 1st edn., Oxford: Clarendon Press. (The 2nd edn., 1994, includes a "Postscript," which is a reply to critics.)
1963: *Law, Liberty and Morality*, London: Oxford University Press.
1968: *Punishment and Responsibility, Essays in the Philosophy of Law*, Oxford: Clarendon Press.
1982: *Essays on Bentham*, Oxford: Clarendon Press.
1983: *Essays in Jurisprudence and Philosophy*, Oxford: Clarendon Press.

Works by other authors

Austin, J. (1954) *The Province of Jurisprudence Determined*, London: Weidenfeld and Nicolson, pp. 184–5.

Dworkin, R. (1977) "The Model of Rules I" and "The Model of Rules II," in *Taking Rights Seriously*, Cambridge, MA: Harvard University Press.
Fuller, L. (1958) "Positivism and Fidelity to Law: A Reply to Professor Hart," *Harvard Law Review* 71, p. 630.
MacCormick, N. (1981) *H. L. A. Hart*, Stanford: Stanford University Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

14

C. L. Stevenson (1908–1979)

JAMES DREIER

Stevenson's major contribution to philosophy was his development of emotivism, a theory of ethical language according to which moral judgments do not state any sort of fact, but rather express the moral emotions of the speaker and attempt to influence others.

Stevenson's emotive theory of ethical language

Stevenson always stressed that his work did not include any substantive moral judgments, but rather comprised "analytic ethics," or what is now commonly called "metaethics," the branch of moral theory that is *about* ethics and ethical language.

What do we mean when we say that something is good or bad, or right or wrong? On the face of it, we are describing, attributing to the thing some property, goodness or badness, or rightness or wrongness. What could these properties be? How do we find out about them? Much of philosophical moral theory explores various answers to these questions. Stevenson thought that questions about the nature of moral properties were misplaced. Our moral judgments do not, at least primarily, describe at all. Uttering moral sentences has a different function: to express emotions, and to influence or invite others to share them. All of his main contributions appeared in *Ethics and Language*, 1944, and a collection of papers, *Facts and Values*, 1963.

Distinguish between *expressing* a certain state of mind and *saying that one is in it*. If I say, "Ann Arbor is in Michigan," I express my belief that Ann Arbor is in Michigan, but I do not *say* that I believe such a thing. For what makes what I said true? Not that I really do believe that Ann Arbor is in Michigan; only the fact that Ann Arbor really is in Michigan. Stevenson's theory of ethical language, in a nutshell, was that when I say, "Inequality is bad," I have expressed a certain negative moral attitude toward inequality, though I have not said that I have it. It should be clear why Stevenson stressed that his theory was "analytic" or metaethical, and did not contain any substantive moral judgments. For by claiming that moral judgments serve to express emotions, he had not expressed his own moral emotions at all.

Besides expressing the speaker's attitude, Stevenson said, moral statements also "create an influence," they invite the audience to share in the emotion expressed. Thus, "*x* is good" is akin to "Let us approve of *x*." Moral exhortation, after all, is commonly

used to try to persuade the audience to share the speaker's suggestions, and moral judgment is often a call to action. Furthermore, in context, ethical statements can come to have some secondary descriptive content; in Victorian England, for example, calling a woman "virtuous" implied that she was chaste. So a Victorian moralist could manage to describe a woman, and not merely to evaluate her (express his emotional attitude toward her and invite others to share it), by calling her "virtuous."

Some advantages of emotivism

Stevenson's theory was enormously influential in the middle of the twentieth century. Taking its cue from Ayer's short remarks on ethics in *Language, Truth and Logic*, Stevenson's theory added sophistication and subtlety (see AYER).

In "The Emotive Meaning of Ethical Terms," Stevenson sets out some criteria for a successful analysis of moral terms, explaining that what he calls traditional "interest theories" of ethical terms fail one or more criteria. These interest theories include the views of Hobbes, whom Stevenson understood to have defined "good" to mean "desired by me," and Hume, whom he interpreted as defining it to mean "desired by my community."

> In the first place, we must be able sensibly to *disagree* about whether something is "good." This condition rules out Hobbes's definition. For consider the following argument: "This is good." "That isn't so; it's not good." As translated by Hobbes, this becomes: "I desire this." "That isn't so, for I don't . . ."
>
> In the second place, "goodness" must have, so to speak, a magnetism. A person who recognizes X to be "good" must *ipso facto* acquire a stronger tendency to act in its favour than he otherwise would have had. This rules out the Humean type of definition. For according to Hume, to recognize that something is "good" is simply to recognize that the majority approve of it. Clearly, a man may see that the majority approve of X without having, himself, a stronger tendency to favour it . . .
>
> In the third place, the "goodness" of anything must not be verifiable solely by use of the scientific method. "Ethics must not be psychology." This restriction rules out all of the traditional interest theories, without exception.

Emotivism appears to be well prepared to satisfy these criteria. First, people can genuinely disagree when one states that X is good and the other states that X is not good; they are disagreeing in attitude, as Stevenson puts it, not in factual belief, but this is genuine disagreement just as plainly as we may disagree when I suggest that we go to the movies tonight and you suggest that we go have a few drinks instead. Second, and perhaps most significantly, if the judgment that X is good is an expression of favorable attitude toward X, then it is clear why anyone making such a judgment will have a tendency to act in favor of X. Finally, while adding to our knowledge by scientific investigation may sometimes resolve certain ethical issues, there can be deeper disagreements that are left untouched by scientific methods. Our emotional attitudes may differ in the face of converging empirical knowledge.

A further attraction of emotivism is that it dissolves knotty-looking metaphysical problems of metaethics. Consider the question of whether moral properties are natural properties or some other special sort. G. E. Moore famously argued that moral proper-

ties could not be natural properties, and later John Mackie argued for skepticism about the existence of moral properties on the grounds that they could not be natural ones, and the metaphysics and epistemology of non-natural properties is too spooky (or "queer," as Mackie said) for sober philosophy. Some metaethicists have tried to show how moral properties could be part of the natural world after all, but it is difficult to explain just how our linguistic habits and practices could determine just which natural property moral wrongness could be, given the wide diversity and disagreement in moral values among different people and cultures at different times. Emotivism resolves the issue by denying that moral predicates, like "wrong" and "good," serve to pick out properties at all. They serve as markers of mood or emotion instead. So the metaphysics of alleged moral properties is avoided if we adopt Stevenson's view.

Some difficulties for emotivism

Emotivism is not without its difficulties, and the main ones were leveled at Stevenson soon after he began to publish his views. One criticism was offered by Brand Blanshard in a paper called "The New Subjectivism in Ethics." Blanshard complained that emotivism has an obviously false implication. When I see a rabbit with its foot caught in a trap, I might say (or think) "That's a bad thing." I would then, plausibly, be expressing my negative emotion toward the pain of the rabbit. But suppose I then contemplate the situation in which I myself become very jaded and cease to care about the suffering of sentient animals. Do I (now, actually) say, "Well, in that case, the suffering of the rabbit would not be a bad thing at all"? No, of course not. But emotivism implies that this is how I should think. So emotivism is false.

This criticism is instructive, though it is not correct. It illustrates two important points about Stevenson's theory. First, the fact that we would *ordinarily say* one thing or another is very important, according to Stevenson's approach. He would never have replied, "We might not say such a thing in that case, we might steadfastly deny it, but we would be mistaken." His theory was supposed to account for our ordinary judgments, and not to reform those judgments. So it is important whether Blanshard's example really does show that emotivism sometimes contradicts our ordinary ethical judgments.

However, the criticism is unsound, because Stevenson's approach does not, in fact, imply that we do or should judge that the suffering of the rabbit would not be at all bad if we were jaded and uncaring. To think that it does imply such a thing is to mistake emotivism for a poor relation, subjectivism. The subjectivist thinks that "bad" means (something like) "apt to cause a negative emotion in me." So to call something bad, according to subjectivism, is to *say that* it causes a negative emotion in oneself. But Stevenson took great pains to distinguish his own view from subjectivism, and he gave very similar examples to show why subjectivism is incorrect. According to emotivism, remember, calling something bad is not saying *that* it does or doesn't do anything – that would be to describe the thing. Ethical language does not (primarily) describe a thing or an emotion or the speaker, it *expresses* the emotion of the speaker. When I contemplate the situation in which I heartlessly feel no sorrow over the rabbit's suffering, I (right now, actually) feel rather bad about that, and if I were to express my emotion I would say, "That would be a bad thing."

Probably the most influential criticism of Stevenson, the criticism that later emotivists (and fellow non-descriptivists, see below) have been most concerned to address, was a problem noticed by Peter Geach and John Searle. It is sometimes called the "embedding problem." To put it succinctly, the problem is that even if emotivism really does tell us what somebody does when she asserts a simple moral sentence like "It is wrong to kick cats," it does not seem to tell us what such a sentence means. For there is more to the meaning of a sentence than the facts about what is accomplished or expressed by an assertion of it, since we can use sentences without asserting them, in *unasserted* or *embedded* contexts. There are many kinds of unasserted contexts. Here are a few examples; notice that in no case would someone sincerely uttering the entire sentence be asserting that it is wrong to kick cats.

If it is wrong to kick cats, then it is wrong to kick Tibbles.
Either it is wrong to kick cats, or there is nothing wrong with kicking people.
I wonder whether it is wrong to kick cats.
Do you mean to say that it is wrong to kick cats?

Many other kinds of examples could be given, but the idea is clear enough. Critics of emotivism point out that what Stevenson said about the emotive meaning of ethical terms does not seem to explain how a sentence like "It is wrong to kick cats" *embeds* into these complex contexts. What, that is to say, does the sentence contribute to the complex whole? One thing is clear enough: someone uttering any of the four example sentences above could not be said to be expressing a negative emotion toward kicking cats. So something more must be said. Stevenson himself never seems to have taken this problem to heart, so he never said much of anything by way of reply. But some later non-descriptivists have said more (see below).

The embedding problem may appear to be a kind of technicality, and perhaps it is, though many philosophers have taken it very seriously. The final criticism I will mention seems to cut deeper into the spirit of emotivism. Stevenson said, and emotivism gains much of its plausibility from this idea, that a person who sincerely asserts or believes a moral judgment must necessarily feel some sort of emotional tug, so that whoever judges something good must be emotively in favor it, and whoever judges something bad must be against it or inclined to avoid, or would try to eliminate it. But we may wonder whether this claim is true. Isn't it possible to judge sincerely that something is good, but feel no sympathy or other "pro-attitude" toward it whatsoever? There is no uncontroversial answer. Some find it obvious that such a thing is possible, while others are at least at first inclined to wonder what the questioner could possibly have in mind. But if we tell a background story it starts to seem very plausible that Stevenson may have overstated the connection between moral judgment and emotion.

Surely it is imaginable that someone could be a self-avowed and sincere amoralist. Such a person would have no interest at all in moral values or rules, and might even be perfectly forthright in admitting so. Yet amoralists could surely learn to recognize which things are good and bad, even if the normal concern with such things might seem quaint or misguided to them. So they could with perfect sincerity and understanding manage to judge that giving to charity is morally good, or that breaking promises is bad, and they could make those judgments without any emotion or

motivation or tendency to promote the "good" things or discourage the "bad" ones. All of this seems possible. Doesn't it show that emotivism is a mistaken theory?

Perhaps not. In the paragraph above, the words "good" and "bad" are in quotation marks. It is plausible that amoralists use these and other moral words in what R. M. Hare called the "inverted commas sense," really *mentioning* them rather than using them. Amoralists cannot say (sincerely, at least) that charity is good, so instead they say that charity is *what most folk call "good."* We (moralists or amoralists) can certainly mention emotive words without expressing their emotive meanings.

Some related theories

Hare's theory is not emotivist, though it is a close ally. According to Hare, the main function served by moral judgments is prescription (see HARE). So Hare agrees with Stevenson that we do not fundamentally describe things when we call them good or bad, and he even agrees that moral judgments could be called "expressions of emotion," since prescriptions are expressions, in a sense. But Hare cautions against taking Stevenson's idea too literally. In *The Language of Morals*, he writes:

> We speak of expressing statements, opinions, beliefs, mathematical relations, and so on; and if it is in one of these senses that the word is used, the theory, though it tells us little, is harmless enough. But unfortunately it is also used in ways which are unlike these; and Ayer's use (in speaking of moral judgements) of the word "evince" as its rough synonym was dangerous. Artists and composers and poets are said to express their own and our feelings; oaths are said to express anger; and dancing on the table may express joy. Thus to say that imperatives [or moral judgments] express wishes may lead the unwary to suppose that what happens when we use one, is this: we have welling up inside us a kind of longing, to which, when the pressure gets too great for us to bear, we give vent by saying an imperative [or moral] sentence. Such an interpretation, when applied to such sentences as "Supply and fit to door mortise dead latch and plastic knob furniture", is unplausible.

In the 1980s and 1990s Simon Blackburn and Allan Gibbard developed versions of emotivism (or in Gibbard's more general terminology, "expressivism") grounded in the same root ideas as Stevenson's theory. These theories are more sophisticated in various ways (in particular they make good headway into the embedding problem mentioned above), and they have to some extent supplanted Stevenson's emotivism, though as inheritors, not as refuters.

Bibliography

Works by Stevenson

1944: *Ethics and Language*, New Haven, CT: Yale University Press.

1963: *Facts and Values: Studies in Ethical Analysis*, New Haven, CT: Yale University Press. (A collection of papers. Essay Two is especially useful as an introduction. Essay Eleven contains the most mature version of the theory.)

Works by other authors

Ayer, A. J. (1936) *Language, Truth and Logic*, New York: Oxford University Press. (Chapter 6 contains the germ of emotivism, which Stevenson developed into a full and sophisticated theory.)

Blackburn, S. (1984) *Spreading the Word*, Oxford: Oxford University Press. (Chapter 6 especially develops a variation on emotivism designed to address prominent objections.)

Blanshard, B. (1949) "The New Subjectivism in Ethics," *Philosophy and Phenomenological Research* 9, pp. 504–11.

Geach, P. T. (1960) "Ascriptivism," *Philosophical Review* 69, pp. 221–5. (An influential, but somewhat technical, objection to emotivism.)

Gibbard, A. (1990) *Wise Choices, Apt Feelings*, Cambridge, MA: Harvard University Press.

Goldman, A. and Kim, J. (eds.) (1978) *Values and Morals: Essays in Honor of William Frankena, Charles Stevenson, and Richard Brandt*, Dordrecht and Boston: Reidel. (A collection of critical essays; those on Stevenson's work are quite accessible. Contains a comprehensive bibliography of Stevenson's writing.)

Hare, R. M. (1952) *The Language of Morals*, Oxford: Oxford University Press. (A work contemporary with Stevenson's, with important similarities and contrasts.)

Searle, J. (1969) *Speech Acts*, Cambridge: Cambridge University Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

15

W. V. Quine (1908–2000)

PETER HYLTON

Willard Van Orman Quine was born on June 25, 1908. He was graduated from Oberlin College with a degree in mathematics, *summa cum laude*, in 1930; his senior honors thesis was a proof within the system of Whitehead and Russell's *Principia Mathematica* (1910–13), which he studied largely without aid from his teachers. Whitehead was in the Philosophy Department at Harvard, so it was there that Quine went to do graduate work, although Whitehead was no longer teaching logic and Quine had done little undergraduate work in philosophy. Nonetheless, he completed a Ph.D. in two years, graduating in 1932. His dissertation generalized *Principia*'s treatment of classes so that it included dyadic relations, instead of treating the latter separately. Along the way, Quine clarified and reformulated the basis of the system – a point to which we shall return.

Quine spent the academic year 1932–3 in Europe on a Sheldon Fellowship. He spent five months in Vienna, attending some meetings of the Vienna Circle. More important, perhaps, was a shorter stay in Prague, where he had extensive conversations with Rudolf Carnap, then completing *The Logical Syntax of Language* (Carnap 1934). While in Europe he was elected as one of the first group of Junior Fellows in Harvard's newly formed Society of Fellows: a three-year fellowship, without teaching obligations. He spent most of these years working on logic and set theory, though some of it on other aspects of philosophy. Under this latter head, he gave three lectures on Carnap, essentially expounding, in a strongly approving fashion, what he took to be the doctrines of *The Logical Syntax of Language*. (The text of these lectures is now published in Creath 1990.) In 1936 Quine became Faculty Instructor at Harvard. Except for service in the US Navy during World War II, he held faculty positions there from that time until his retirement in 1978. He remained philosophically active and engaged for twenty years after retirement, continuing to write and publish into his nineties.

It will be helpful to put Quine's work in the context of what I shall call twentieth-century scientific philosophy, a movement within the broader stream of twentieth-century analytic philosophy. Key figures in twentieth-century scientific philosophy (other than Quine) include Bertrand Russell and Rudolf Carnap, as well as others often identified as logical positivists or logical empiricists; Frege and Wittgenstein also made crucial contributions to the movement.

Let us try briefly to characterize this movement by aims and doctrines rather than by its participants. Perhaps most notable is the emphasis on knowledge, and its objects, rather than on ethics or politics or aesthetics or history or the human condition, as the primary concern of philosophy; an emphasis, one might say, on the True rather than on the Good or the Beautiful. This emphasis is equally an emphasis on science, especially on the natural sciences. It is characteristic of scientific philosophy to take the natural sciences as paradigmatic of all knowledge. Part of this view is the doctrine that the Vienna Circle called "the Unity of Science." The point here is the unity of all real knowledge, for the German word *Wissenschaft* is broader in its scope than most current uses of its invariable English translation, "science." (Quine, however, makes it explicit that he uses the word "science" broadly; see Quine 1995: 49.) According to this view, there are no fundamental divisions of aim or method among the various branches of knowledge. Along with this, there is a suspicion, or worse, of the claims of metaphysics, and of any claims neither answerable in straightforward fashion to the findings of empirical science nor provable by logic and mathematics.

This brief sketch at once raises questions about there being any role for philosophy, even that of the scientific philosophers themselves. A central idea here was that philosophy is not to add to our knowledge but is, rather, to *analyze* the knowledge that the sciences give us, and thereby to give us greater clarity about that knowledge and its basis. The tool of this analysis was, above all, logic: the logic of Frege and Russell (see FREGE and RUSSELL). This logic held out an ideal of clarity; one aspect of the philosopher's task was to impose a similar clarity upon other subjects. This line of thought suggests an assimilation of philosophy to logic, but does not by itself account for the possibility of either of these subjects. Russell sought to do this by postulating an a priori insight, which might strike some as a large concession to metaphysics, in the pejorative sense. Carnap, drawing on the early work of Wittgenstein, held that logic is analytic, empty of content, and hence not genuine knowledge at all. Philosophy too makes no claims about the world. Its analyses of language simply make explicit what is already there; it recommends a certain kind of language for this or that scientific purpose, but a recommendation is not a claim, and is presumably not in need of the same sort of justification. This emphasis on language is connected with Carnap's view that analytic truths are true in virtue of language, true by virtue of the meanings of the words making them up. Given the importance of analyticity as accounting for logic, for mathematics, and for philosophy itself, this throws an enormous explanatory burden on the notion of language.

We have just very briefly sketched the tradition of twentieth-century scientific philosophy. Quine's position relative to this tradition is ambivalent. On the one hand, he is its greatest exponent in the last forty years of the century. On the other hand he revolutionizes it, in such a way that one might say that he rejects the tradition rather than continuing it. Both Russell and Carnap attributed great importance to the natural sciences but nevertheless held that logic, mathematics, and philosophy itself, all have a status that is quite different from that of, say, physics or chemistry, or history or sociology. The former are independent of observation, and thus a priori – however exactly that idea is to be understood – while the latter are a posteriori, empirical and ultimately answerable to observation and sensory experience. Quine rejects the

idea that there is a fundamental epistemological distinction here. This rejection – which Quine himself sometimes speaks of as his *naturalism* – is fundamental for his philosophy. We need to see why he rejects the a priori, and how he can get by without it; we shall then to begin to show how his general approach to philosophy flows, in large part, from this step.

Analyticity and the a priori

Like Carnap, Quine rejects any idea of the a priori as based on pure intuition, or on pure reason; such an idea runs counter to his scientific and empiricist predilections. Carnap appealed to the idea of analyticity as an alternative (see CARNAP). Quine, famously, also rejects Carnap's use of this idea and with it any significant idea of the a priori or of necessity. We shall discuss Quine's arguments against Carnap's notion of analyticity, or against the idea that there is a serious and significant distinction to be made between the analytic and the synthetic. Quine's rejection of the distinction, however, is only half the story. The other half is to show how he can make sense of the apparently a priori status of logic and mathematics without it – or, better, perhaps, how he can account for those facts which have led philosophers to think that logic and mathematics must be a priori.

Understanding Quine's attack on Carnap's notion of analyticity is complicated, partly because it was for a time, I think, not entirely clear even to Quine himself exactly what he is attacking and how. Quine thinks of Carnapian analyticity as truth in virtue of meaning, and so also thinks that if we had a clear understanding of the notion of meaning – more precisely, of synonymy, or sameness of meaning – then we would have gone a long way towards making clear sense of a notion of analyticity. So for a long time Quine's attack on analyticity seemed to be part and parcel of an attack on the notion of meaning, as unclear or undefined. And certainly Quine is skeptical as to how far we can make clear sense – which for him means sense in scientific, especially behavioral, terms – of the idea of synonymy. (He is not, however, skeptical of the notion of meaning*fulness*. See his essay "The Problem of Meaning in Linguistics," in Quine 1961, the burden of which is precisely that the notion of meaningfulness is not afflicted with the same sorts of problems as the notion of sameness of meaning, but lends itself to a ready, if somewhat rough, understanding in behavioral terms.) His view of synonymy is *not* that there is no sense at all to be made of it anywhere, even though he has reason to think that we may not be able to make complete sense of it everywhere. So we see Quine, as early as "Carnap and Logical Truth" (1963, written in 1954) accepting that there may be a limited notion of analyticity to be had. His willingness to accept some notion of analyticity becomes more marked as time goes by. In *Roots of Reference* (1974) he proposes a tentative definition of synonymy, and with it an understanding of analyticity; in "Two Dogmas in Retrospect" (1991) we find Quine arguing that (first-order) logic is analytic. How are we to understand this situation?

Quine continues to reject the idea of a notion of analyticity that would play anything like the central philosophical role that Carnap allotted it. In order to play that role, a notion of analyticity would have to meet two requirements. First, it would have to have the right scope: the truths of logic and mathematics, at least, must come out as analytic. Second, it must also, at least in Quine's view, mark a significant epistemo-

logical distinction: analytic truths need no justification (or else what counts as "justification" for them is wholly different in kind from the justification of synthetic truths). While Quine accepts *a* notion of analyticity, it is not one that satisfies either of these requirements. Let us begin with the question of scope and definition.

Quine, as we saw, takes analyticity to be truth in virtue of meaning. But how are we to understand the idea of meaning, as it occurs here? For Quine, the only thing that could be relevant to the meaning of a word or a sentence in a given language is how it is used by speakers of that language. This is an important point. Quine has been accused of being unduly behavioristic, especially about language. Certainly he has a general bias in favour of a behaviorist approach to the mind. He claims, however, that his insistence on approaching language-use behaviorally is not merely the result of prejudice. Indeed he offers an argument for some form of behaviorism in this context. The passage is worth quoting at some length:

> In psychology one may or may not be a behaviorist but in linguistics one has no choice. Each of us learns his language by observing other people's verbal behavior and having his own faltering attempts observed and reinforced or corrected by others. We depend strictly on overt behavior in observable situations. As long as our command of our language fits all external checkpoints . . . so long all is well. Our mental life between checkpoints is indifferent to our rating as a master of the language. There is nothing in linguistic meaning beyond what is to be gleaned from overt behavior in observable circumstances. (Quine 1990: 37–8)

For Quine there can be no more to meaning than is implicit in the actual use that is made of the language.

Quine's interest is exclusively in knowledge, and the aspect of the use of language that primarily concerns him is our accepting or not accepting sentences. Thus he says, early on: "in point of *meaning* . . . a word may be said to be determined to whatever extent the truth or falsehood of its contexts is determined" (1936: 89). But then the question is: *which* of the contexts of a word must be so determined in order to determine its meaning? Without some reason to discriminate, we have no reason to treat one context as more definitive of a word's meaning than any other. But then no true sentence in which the word appears would have any better claim to be analytic than any other such sentence; clearly no useful analytic/synthetic distinction can be erected on that basis.

What sort of thing might give us reason to discriminate among contexts? If mastery of some small subset of a word's uses gave one mastery of its use as a whole, then there would be reason to say that those uses, those contexts, constituted its meaning. And clearly this happens in some cases. A child who otherwise has a fair degree of linguistic sophistication but does not know the word "bachelor" can be given a mastery of that word all at once, at a single stroke, by being told that bachelors are unmarried men. This fact gives us every reason to say that "bachelor" *means* "unmarried man," and that the sentence "All bachelors are unmarried" is analytic – which Quine, at least in his later work, certainly accepts (see Quine 1991: 270). Along these lines, he proposes a definition of analyticity: "a sentence is analytic if *everybody* learns that it is true by learning its words" (1974: 79). He argues that first-order logic is analytic by this sort

of definition, but that other analytic truths will all be trivial. In particular, there is no prospect of arguing on this sort of basis that mathematics is analytic; apart from other considerations, Gödel's incompleteness theorem would be an insurmountable barrier to such an argument (see TARSKI, CHURCH, GÖDEL).

We have yet to discuss the question of the epistemological significance of the notion of analyticity. This, Quine came to see, is the crucial question; in the 1980s he wrote: "I now perceive that the philosophically important question about analyticity and the linguistic doctrine of logical truth is *not* how to explicate them; it is the question of their relevance to epistemology" (Hahn and Schilpp 1998: 207). Why should anyone think that showing a sentence to be analytic for a given language – learned in the course of learning the language – shows anything about its epistemological status? Why might one think that it shows that for that sentence no justification is required, or that the question of justification is somehow misplaced? Well, clearly it might be thought to show that *given that we are speaking that language* the question of the justification of that particular sentence does not arise. But why does the question not simply become one of the justification for speaking that language? We are presumably operating here with very tight identity-criteria for languages, so that shifting the meaning of the one word "bachelor" would mean that we were speaking a different language (if this seems excessively odd, we might speak in terms of idiolects rather than languages; but the point is the same). And given that conception of a language, it is not obviously absurd to ask for the justification for speaking a given language (it is no longer enough to say, "it is the one I was brought up with and feel most at home in," for this quality would survive minor shifts).

For Carnap, the choice of a language is in epistemologically important ways unlike the choice of a theory within a language. The former is not a matter of *correctness*, of right or wrong; it is a practical matter having to do with pragmatic factors such as the simplicity of a given language and its convenience for this or that goal. This idea issues in what he calls the "Principle of Tolerance": since choice of language, unlike the choice of a theory within a language, is not a matter of correctness or incorrectness, we should be tolerant, and allow people to work with whatever language they choose. Within a given language, justification is more or less rule-governed, governed by the rules of that particular language; justification, like other significant philosophical notions in Carnap's view, is thus language-relative. But the choice of a language itself is not something that can be justified in the same sort of way, since without a language we have no rules of justification to which to appeal.

Carnap emphasizes the distinction between the justification of choice of a theory within a language and the justification (or the lack of need for justification) of choice of language. The idea that attributing analyticity to a sentence has epistemological significance depends upon this distinction. Saying of a sentence that it is analytic would mean that it is in some sense integral to the language that we currently speak. So if we ceased to accept that sentence we would have modified the language. But that would leave open the possibility that we might have evidence which would justify that modification of the language. Thus it would seem that evidence might bear on an analytic sentence in the same sort of way in which it bears on a synthetic sentence, unless the notions of evidence and justification in play are of different kinds in the two cases. Carnap, of course, holds there there is just such a difference in kind. He claims that

there are quite different conceptions of evidence and justification at work. Within the language, justification is a rule-governed procedure, and a matter of right or wrong; when the language itself is being chosen, however, there are no rules to which to appeal, and the choice is purpose-relative and to some extent arbitrary.

Quine attacks this distinction from both sides. He denies that (internal) justification is to any significant extent a rule-governed procedure. He says, for example:

> I am impressed . . . apart from prefabricated examples of black and white balls in an urn, with how baffling the problem has always been of arriving at any explicit theory of the empirical confirmation of a synthetic statement. (1961: 41–2)

He also insists that all our cognitive choices, including the choice of a language for knowledge, are directed towards the same end: achieving the most successful theory, where a crucial test of success is the generation of true predictions. Vaguer virtues, such as simplicity and fruitfulness are also relevant. These are the sorts of things that Carnap counted as "pragmatic factors," applicable to questions of language-choice. Quine claims that they are applicable also to what Carnap would count as empirical beliefs. They may not in any very obvious way be applicable to the question whether there is now a desk in front of me, but certainly they are to more or less abstract claims of theoretical physics. For Quine there is a continuum here, with no sharp breaks to be had.

Quine thus holds that even where we have a significant truth which is analytic, this status simply does not matter epistemologically:

> "Momentum is proportional to velocity" counts as analytic. But do we care? Einstein's relativity theory denies the proportionality law, complicating it with a formula involving the speed of light. But instead of accusing Einstein of a contradiction in terms, we simply stand corrected. (Quine 2000)

Now Carnap might agree that we "stand corrected" because we accept that Einstein has shown us that a non-Newtonian language works better for making some predictions, but he would insist that this is a different sense of correction from that in which we are corrected when we change our mind about a belief that does not involve a change of language. But this is precisely what Quine denies, as we have seen.

One issue which arose above was Quine's view of the nature of justification, the relation between the evidence we have and the beliefs that we hold on the basis of it. The point there was that there is not, in general, a simple relationship between a sentence, on the one hand, and an observation or group of observations that justify it, on the other hand. Justification is not, in general, a simple and rule-governed matter. Of course there are sentences, such as "there is a desk in front of me now," which do seem to have a very straightforward relation to observations. What makes that case straightforward is that it hardly matters what else a person believes: given the right observations, almost anyone will accept that there is currently a desk in front of them. The justification relation here holds between observations and the individual sentence believed, whatever one's other beliefs may be. In Quine's view, however, this is a poor paradigm to use for knowledge as a whole. In general, justification is *holistic*, meaning that it does not apply

to sentences taken individually, in isolation from others, but rather to larger or smaller chunks of theory, made up, in some cases, of a large number of sentences. Many of the sentences we accept – most obviously the more abstract and theoretical ones – have relations to observations only if we tacitly assume many other sentences. These other sentences, background assumptions, are required if the sentence in which we are interested is to have any observational consequences at all. From the point of view of the working scientist, the background assumptions may be confidently accepted, and only the individual sentence up for testing. From a more abstract point of view, such as Quine's, however, what is tested by observation is not the individual sentence alone, but rather the whole set of sentences that implies the observational consequences. From a sufficiently abstract point of view, indeed, it is always the whole of our knowledge that is tested. Any test of a sentence presupposes truths of logic among its background assumptions. Logic, however, is used everywhere in our system of beliefs, so in a rather Pickwickian sense it is that system as a whole that is at stake. (Quine calls this extreme holism "legalistic" (1991: 268). I take this to mean that it holds from a very abstract point of view, but not that it is unimportant.)

Holism is not new with Quine. When "Two Dogmas" was reprinted, Quine added a footnote to Duhem, and there is a clear statement of the view in Carnap's *Logical Syntax of Language*, with references to Poincaré as well as to Duhem (Carnap 1937: 318). Quine's uses of the doctrine, however, are novel. One use we have seen: it is the basis of the claim that justification (within a language) is not the sort of rule-governed procedure that Carnap sometimes suggests, and so is not different in kind from the sort of justification that applies to the choice of one language rather than another. This claim, in turn, is crucial for Quine's view of analyticity and the a priori, discussed in the previous few pages.

A second use that Quine makes of holism also relates to the question of the a priori, but in a quite different way. He attempts to undercut the idea that there must be a priori knowledge by invoking holism to explain the phenomena which led some philosophers to invoke the idea of the a priori, but to do so without invoking that idea. (If those phenomena can indeed be explained without the a priori, then their existence no longer constitutes a reason to accept the a priori.) There is no doubt that the theorems of mathematics and of logic are not discovered by experiment or, at least in any ordinary sense, justified by observation. And their falsehood seems completely inconceivable. How can these facts be explained in accordance with Quine's views? Holism provides the answer. Logic figures everywhere in our system of beliefs; mathematics is used in many branches of knowledge. No one observation or experiment bears on them, but the success of our system of beliefs as a whole in predicting experience provides justification; this justification is exceedingly indirect. For Quine that only puts it at one end of a continuum which we already have reason to accept, for "$e = mc^2$" is justified very much less directly than is "there is a desk in front of me now." Equally the unimaginability of the falsity of logic becomes intelligible. Given the ubiquity of logic, changing it means making changes everywhere in our system of knowledge. It is not to be wondered at if this is hard to conceive.

Our discussion so far has been focused on Quine's rejection of the a priori. He is thus left with no kind of knowledge other than the ordinary knowledge of common sense

and (better) science. Philosophy too, since it claims to yield knowledge, must be of this same general sort. This is the doctrine that Quine calls naturalism; it is absolutely fundamental to his thought. The rest of this essay will take naturalism as its starting point, and investigate Quine's philosophy as an unfolding of that doctrine. I shall begin with topics having to do with epistemology, and then move to topics whose focus is ontology and metaphysics (or its Quinean analogue). There is, however, one further issue which we should briefly mention here.

Quine has argued for the possibility that two translators might come up with different translations of some sentences – not merely stylistically different, but "not equivalent in any plausible sense of equivalence, however loose" (Quine 1960: 27). This is the controversial doctrine known as the indeterminacy of translation. Where sentences are not asserted or denied on any very direct observational basis, the evidence (behavioral evidence, of course) for one translation over another is mediated by other sentences; alternative translations of all of them might cancel out, leaving each of two overall schemes of translation as equally justified. In Quine's view there would, in such a case, be no right and wrong, no fact of the matter: it would not be our knowledge which was lacking, but rather that there was no fact to be known. This idea provoked an enormous amount of discussion in the 1960s and early 1970s, and some commentators have even thought that it is what really underlies Quine's objections to analyticity. In my view, however, it is not of great importance to his thought, except as dramatizing the idea that meaning must ultimately be answerable to behavior (see Hylton 1990). Note in this connection that Quine now speaks of indeterminacy as "a conjecture" (see Hahn and Schilpp 1986: 728).

Knowledge and the realm of the cognitive

How should we conceive of knowledge, if we are to take a scientific approach to it? Fundamental to Quine's thought is the idea that knowledge is to be understood as a biological phenomenon. Human knowledge is thought of as a condition of the human animal. It originates in the struggle of one species of primate to survive. The opening sentences of *From Stimulus to Science* read like this: "We and other animals notice what goes on around us. This helps us by suggesting what we might expect and even prevent, and thus fosters survival" (Quine 1995: 1). This is what one might call a Darwinian conception of knowledge: knowledge as an adaptive mechanism, fostering the survival of the species.

In this view we see a decisive rejection of the long-standing philosophical tradition that sharply distinguishes between real knowledge and mere belief or opinion, between *scientia* and *doxa*. That tradition tends to assimilate real knowledge – *scientia* – to knowledge of a proposition of mathematics, when it is known on the basis of a thoroughly understood proof. Real knowledge is accordingly thought of as infallible and known with certainty. None of these ideas fits with the idea of knowledge as a biological phenomenon, as what helps the human animal to get by in its dealings with the world. Some philosophers have held that those ideas are implicit, more or less, in the word "knowledge." Quine's response is simply to abandon that word for "scientific and philosophical purposes" (1987: 109). He continues to use it informally (as shall I in expounding his views), but without the weight that it may be thought to carry in

general use. He makes no sharp distinction between our knowledge, and that which we, or experts among us, accept upon reflection. Quine's conception of cognition, then, is fallibilist through and through: no part of our system of beliefs can be counted as wholly immune from revision, though some parts are no doubt far more secure than others.

A second very general point about Quine's conception of knowledge is that he takes it to be linguistic or verbal; at least for "scientific and philosophical purposes," he thinks of our system of beliefs as being embodied in sentences:

> What sort of thing is a scientific theory? It is an idea, one might naturally say, or a complex of ideas. But the most practical way of coming to grips with ideas, and usually the only way, is by way of the words that express them. What to look for in the way of theories, then, are the sentences that express them. (Quine 1981: 24)

Much of Quine's interest in language and in its analysis arises from the fact that our knowledge is embodied in language.

In one way, as we saw, the idea of knowledge, as Quine employs it, marks no very sharp distinction. Unlike some philosophers, he does not use the term as an honorific, connoting some particular high degree of justification or of certainty. He does, however, make a sharp distinction between the realm of the cognitive and the rest of human activity. To call a human activity cognitive is, roughly, to say that it is answerable to, if not exclusively aiming at, predictions of sensory experience. It is perhaps a presupposition here that cognitive activity and cognitive language can be peeled off from the chaotic mass of human activity and language generally – or at least that we can abstract without distortion, and talk of the cognitive while ignoring the rest.

The prediction of experience is a practical matter; at the limit, as we indicated, survival is at stake. This is the sense in which Quine's conception of knowledge is Darwinian. Thus far Quine is with the pragmatists. On the other hand, tying the concept of knowledge to the prediction of sense-experience enables Quine to make clear distinctions, and to erect barriers, in places where the pragmatists would not. Activities which predict experience are cognitive; others, though they may contribute to human flourishing in other ways, are not. The justification for this distinction is presumably (for Quine is not explicit here) that sense-experience is our only way of finding out about the world (we shall discuss this idea shortly). In spite of Quine's practical, Darwinian view of knowledge, there is thus a sense in which his view makes a clear distinction between the theoretical and the practical. *Theoretical* success is success in prediction of sensory experience.

Evidence

As traditionally conceived by philosophers, this notion includes two strands. On the one hand, evidence is thought of as consisting of the epistemologically most fundamental items of our knowledge: evidence is that which is, so to speak, first in the order of knowledge (that which is evident), and so also that from which other items of knowledge must be inferred. On the other hand, evidence is also to consist of immediately given data, devoid of any conceptual impositions of our own; since no interpretation is

involved, there is no room for doubt. These two strands are in tension. What is literally first in the order of knowledge seem to be facts about other people and ordinary physical objects. Yet sentences recording such facts do not seem simply to record raw data. They involve conceptualization, and are (notoriously) open to doubt.

Quine's response to these difficulties is to abandon the traditional conception of evidence completely, in favour of a physicalistic alternative. He speaks not of "the given" but of the stimulation of our sensory receptors, and of observation sentences: roughly, sentences that any speaker of the language is disposed to accept or reject simply on the basis of current stimulation. Thus he says:

> our immediate input from the external world [is] the triggering of our sensory receptors. I have cut through all this [i.e. the difficulties of analyzing the notions of observation and experience] by settling for the triggering or stimulation itself and hence speaking, oddly perhaps, of the prediction of stimulation . . .
>
> Observation drops out as a technical notion. So does evidence, if that was observation. We can deal with the question of evidence for science without the help of "evidence" as a technical term. We can make do instead with the notion of observation sentence. (1990: 2)

What is at stake in philosophical talk of evidence is, from Quine's point of view, the issue of how we find out about the world. Here we have a crucial example of his method, of the idea that the study of knowledge is to be naturalized, and as far as possible put on a physicalistic basis. For him, the issue is to be taken as a scientific question: how do we come by information about the world? And the answer is that we do so by the impact of various forms of energy on our sensory surfaces. Physics will tell us what forms of energy there are; physiology and psychology will say which forms of energy human beings can detect, i.e. to which forms human beings respond. The central fact here is that it is only through stimulation of our nerve-endings by energy impinging on our sensory surfaces that we human beings know anything at all about the world. This is fundamental to Quine's epistemology and it is, he emphasizes, "a finding of natural science itself" (1990: 19).

Quine's use of the notion of stimulations is thus symptomatic of his general shift in perspective. The question is not: what is given to me, at the outset of my cognitive endeavors? But rather: how do we humans gain knowledge of the world? This is on his view a straightforward scientific and causal question. The answer refers us to stimulations of our sensory surfaces. Quine thus abandons, or greatly modifies, the traditional concept of sensory evidence. In so doing he shifts the question to which the original conception was an answer. The traditional philosopher's demands are not met by Quine's view. In the most straightforward sense, the occurrence of stimulations is independent of and prior to theory. Our *knowledge* of such matters, however, is clearly not independent of theory. So there is here no prospect of what the traditional philosopher sought: support of the theory which is wholly independent of the theory. To the extent that this strikes us a problem or a paradox, to that extent we have not accepted Quinean naturalism.

The point we have just noted links Quine's rejection of the sensory given with his rejection of the traditional conception of the a priori, or of any analogue. The given

and the a priori were each conceived, by some philosophers, as a kind of extra-theoretical knowledge, knowledge somehow free of the vicissitudes affecting the ordinary knowledge of common sense and science. Quine denies that there is any such kind of knowledge. Within what we take ourselves to know there is, no doubt, better and worse; there is, however, no knowledge of a wholly different and superior kind. Quine is consistent here. He does not take even the most fundamental points of his own philosophy to be a priori – and this includes the doctrine that we know about the external world through impacts on our sensory surfaces. While it is extremely unlikely, there are imaginable circumstances under which we might drop our present idea of sensory evidence entirely. If sufficient confusion resulted from our following the evidence of our senses, as we now understand them, and sufficient success on the part of those who claim to hear voices in their heads, say, our estimate of the role of stimulation of our sensory surfaces might change. (This point is explicit; see Quine 1990: 20f.)

For Quine, as we have said, the central fact about knowledge is that it is only through stimulation of our nerve-endings that we know anything at all about the world. Such stimulations provide the only empirical constraint on our system of beliefs, the only external criterion of success. (By speaking of an *external* criterion I mean to leave room for what one might think of as internal factors: such as the overall simplicity of the system.) This fact suggests that there may be empirical slack between evidence, even the totality of all possible evidence, and theory. We can focus this idea by asking whether there might be two systems of belief, different from one another but each fully successful at "predicting stimulations." Quine's answer, though somewhat qualified and complicated, is that nothing rules this out. This is the doctrine known as *the underdetermination of theory by evidence*. Even if, *per impossibile*, all the evidence were in, still any given theory based on that evidence might not be uniquely justified. If we had a theory that explained and predicted the evidence satisfactorily, that would of course justify it; since it would at least in principle be possible for another theory to do as well, however, the justification would not be unique.

The relation of evidence to knowledge: observation sentences

So far we have said something about evidence and something about knowledge, as Quine conceives them, but nothing explicit about the relation between the two. Clearly there is a gulf between the sensory stimulations that are our only source of information about the world, and the mass of sentences in which our beliefs about the world are embodied. How is this gulf bridged? For the sentences most directly tied to sensory evidence, *observation sentences*, there is a fairly clear account. The other sentences of our theory of the world get their relation to evidence *via* their relation to observation sentences, and here the account is much sketchier. (A full account is, Quine thinks, not yet available, and may never be.) Discussing observation sentences will at least give us an idea of how the gap between evidence and theory is bridged.

The first point to make about what Quine calls observation sentences is that they are – by his lights – *sentences*. This does not mean that they are all sentences in the grammatical sense: "Rabbit," taken as a complete utterance, is Quine's own example. The idea of a sentence here is that of a piece of language which may be used to

say something, and thus may be true or false. For Quine, the fundamental evidential relation holds between sensory stimulations and units of language of that sort.

We can bring out the significance of this idea by contrasting sentences (as we are using that word) with referring expressions, or "terms," as Quine says. In most uses, the word "rabbit" is a term, not a sentence: if I say "I see a rabbit" then in that sentence it is a term which functions as part, but only part, of a sentence. Mastery of the word in that sort of use, Quine claims, requires more than a mere ability to respond to the presence of rabbits. It requires also that one can distinguish the circumstances that license the claim "There's one rabbit" from those licensing the claim "There are many rabbits"; that one can similarly distinguish "There's the same rabbit again" from "There's another rabbit"; and so on. The ability to make these and related distinctions requires some knowledge of the ways of physical objects in general, and of rabbits in particular. Sentences containing terms are thus not epistemologically basic. Mastery of the use of a term already requires that one possess some knowledge, and then the question of the evidence for that knowledge must arise. Mastery of the word "rabbit" as an observation sentence, by contrast, requires only the capacity to respond to environments containing rabbits in ways in which one does not respond to environments which do not contain rabbits; this capacity requires no auxiliary knowledge. This kind of ability is all that is presupposed by the mastery of an observation sentence. It is primitive and fundamental. It is what underlies all cognitive language-use, including the most sophisticated, but other forms of language go beyond it in principle.

To this point we have said little more about observation sentences than that they are sentences, though that idea has proved to be far from trivial. Beyond that, an observation sentence is what Quine calls an "occasion sentence," i.e. it is one that is true when uttered under some circumstances and false under others (so "It's raining in London" is an occasion sentence; "Gold is a metal" is not). The rough idea that Quine intends to capture is that an observation sentence is an occasion sentence about which there is community-wide agreement under any given circumstances. A little more precisely, we can distinguish two conditions which an occasion sentence must satisfy to count as an observation sentence. First, whether a given speaker of the language is disposed to assent to, or dissent from, an observation sentence at a given time is simply a matter of the stimulations that that individual is undergoing at that time. For each individual, the same stimulation-pattern will typically lead to the same verdict each time. Second, any fully competent speakers of the same language, in the same circumstances, will agree on an observation sentence; Quine speaks of "unhesitating concurrence by all qualified witnesses" (1995: 44). (There is some vagueness here, arising from the vagueness of "in the same circumstances," and of "qualified witness.")

Observation sentences thus assert the presence of something readily detectable by the senses. Whether such a sentence is correctly assertable in a given situation does not depend upon ancillary information, unless it is shared by all speakers of the language. Hence "It's cold here!" might qualify, but "That's Quine!" would not, since not all English-speakers would recognize that philosopher on sight. The range of observation sentences will vary with our decision as to exactly who should be included among the fully-functioning speakers of the language. "That's red" will presumably count if we exclude the blind and the color-blind, but otherwise not. Observation sentences are the epistemologically most basic parts of our theory of the world. They can be known before

anything else. (They thus play one of the roles traditionally accorded to the notion of evidence: they are first in the order of knowledge.) The knowledge that they embody is so rudimentary that in almost all cases it goes unspoken, but can be elicited by raising the sentence as a question and noting the subject's reaction.

Considered holophrastically, i.e. as unanalyzed wholes, observation sentences are simply responses to stimulation, and are in only the most minimal sense conceptual or theoretical. Hence such sentences can be mastered by a child otherwise quite innocent of language. As Quine says, "Their direct association with current stimulation is essential if the child is to acquire them without prior language" (1990: 5). Considered as made up of parts, however, they connect with sophisticated theory, for the words which make them up recur in more theoretical contexts. This dual aspect is essential to the function of observation sentences as the starting point of language and conceptualization. Because they presuppose so little, observation sentences will be the first sentences learned by a child (or, indeed, by an adult trying to find his or her unaided way in a wholly strange linguistic community); their learning presupposes no prior conceptual or theoretical resources. Because their terms recur in higher theory, learning such a sentence is a start on learning the language as a whole; it is only this sharing of vocabulary which unites the observation sentences with the rest of the language (1990: 8).

Observation sentences, we saw, are occasion sentences: the truth-value of such a sentence will vary from one occasion of utterance to another. Our scientific theories, however, and most other serious knowledge, consists of standing sentences: sentences true or false once-for-all. How is this gap in turn to be bridged? Quine's answer appeals to the notion of an *observation categorical*. This is a sentence compounded of two observation sentences, saying that whenever one of them holds the other will also hold: "Whenever there is smoke there will be fire," for example. This is a standing sentence, and so might be implied by a serious branch of organized knowledge. On the other hand, both of its component parts are observational (or so we are supposing). So we can tell right off whether a given situation is one in which there is smoke, and whether it is one in which there is fire. Hence we can tell right off whether a given situation is one in which the observation categorical is falsified. Because the sentence is in effect a generalization over all situations, it cannot, of course, be *verified* by a single situation, but it – and hence the theory which implies it – can be falsified. Quine readily accepts this asymmetry between verification and falsification, which fits with his general fallibilism. Since our theories have infinitely many observational consequences, they cannot be conclusively verified; in principle, however, a single observation may falsify a theory.

Naturalized epistemology and normativity

We have been articulating Quine's general conception of knowledge, evidence, and the relation between them. This conception is thoroughly naturalistic: Quine treats knowledge as a natural phenomenon, to be studied by the procedures of science. Most of Quine's own work in epistemology is an articulation and defense of this very general conception. He suggests, however, that there is also room in epistemology for detailed piecemeal work of a more recognizably scientific kind. This would consist in tracing the

connections between theory and evidence in a psychologically realistic fashion, to see how our knowledge is in fact related to the evidence that we have. Epistemology of this sort is thus a branch of psychology. (Of course psychology is itself among the items of knowledge whose relation to evidence is to be investigated in this way; Quine speaks here of "reciprocal containment" (1969: 83).)

Epistemology, as traditionally thought of, is a normative subject: it aims to tell us not merely about what is but also about what ought to be; it aims to tell us not only what we do in fact believe, and on what evidence, but also which beliefs are justified on the basis of the evidence that we have. Quinean epistemology, at least according to his account of the matter, is descriptive. To what extent, if any, can this descriptive subject take on the burden of traditional normative epistemology? This is a large and complex question (see Gregory 1999, to which I am indebted here). Roughly we may say that Quine has no room for the very large-scale questions and doubts which are one kind of starting point for traditional epistemology. He has no sympathy at all, for example, with global skepticism. The aim of our knowledge is to predict sense-experience. A theory that does that satisfactorily does all that we can ask. There is no further question as to whether it tells us about the nature of reality:

> what if . . . we have achieved a theory that is conformable to every possible observation, past and future? In what sense could the world then be said to deviate from what the theory claims? Clearly in none. (Quine 1981: 22)

We cannot divorce the idea of reality from that sense-experience. (This point will emerge further in the next section, below.)

There is, nevertheless, a sense in which Quinean epistemology is normative. The criterion of success for all putative knowledge, for science in Quine's broad sense, is the prediction of sense-experience. Quine sees this as defining the notion of science (1990: 20). This definition is not arbitrary: our primary aim, in science, is to find out about the world, and one thing we know – a well-established piece of scientific knowledge – is that it is only through sense-experience that we come to know about the world. Given that our goal is fixed, there are questions of a normative sort about the best ways in which to achieve that aim, and in this instrumental sense epistemology is normative:

> Naturalization of epistemology does not jettison the normative and settle for the indiscriminate description of ongoing procedures. For me normative epistemology is a branch of engineering. . . . There is no question here of ultimate value, as in morals; it is a question of efficacy for an ulterior end, truth or prediction. The normative here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed. (Hahn and Schilpp 1998: 664–5)

Realism

As we have seen, Quine's naturalism can be identified with the view that there is essentially only one kind of knowledge. In particular, there is no special philosophical perspective from which we can attain knowledge that is independent of our

ordinary scientific or commonsensical theory of the world. Thus on his account we are always inside that theory, modifying it, perhaps, but not wholly transcending it. There is no transcendental standpoint that is independent of our ordinary knowledge, and from which we can evaluate that knowledge without presupposing it. Quine's realism is an important application, and illustration, of this view.

Quine's work is full of remarks which might suggest that he does not take our theories – including the "theory" that is common sense knowledge – to be (really) true, and does not take the objects that those theories presuppose to be (really) real. He says that our theories far outrun the evidence that we have for them, and that more than one theory is compatible with that evidence. The disparity between our evidence and our knowledge is a recurrent theme in his work. His insistence on this disparity, together with his view that our knowledge is justified by its efficacy in predicting and understanding the course of experience, might lead one to suppose that his is an instrumentalist or pragmatist view: that our theories, even at their best, are not really *true*, that they do not aim to correspond to an extra-theoretical world, but are simply useful instruments for predicting, understanding, and controlling future experience. Some critics have taken this to be Quine's view, and have seen his insistence on realism as a contradiction, or as a mis-statement of his actual position (see, for example, Lee 1986, and Smart 1969). Quine, however, insists that his view is "robust realism" (1990: 21), and that appearances to the contrary can be dispelled by taking naturalism seriously enough.

On Quine's view, the objects that our theories presuppose *do* exist in extra-theoretic reality. It is part of our theory – that is, part of the best understanding of the world we have – that those objects (with a few exceptions, most of them straightforward) are not dependent on us or our theorizing. Now the critic may protest: it may be part of our theory that our theoretical objects really exist, independent of our theory – but that's just part of the theory. Do the objects *really* exist? But this is an attempt to ask a question from a stance independent of our theorizing about the world: the point of Quine's naturalism is that there is no such stance. The objects that we believe in exist, and are real, in the only sense of those ideas that we actually have.

More generally, the apparently skeptical remarks that Quine makes when discussing our acquisition of knowledge do not affect his belief in the truth of the knowledge that is thus acquired: his ontology is in this respect insulated from his epistemology. This is a consequence of Quine's version of naturalism, and in particular of the reciprocal containment of science, with its ontological claims, within epistemology, and vice versa. Let us see how this goes. We accept, let us suppose, the best overall theory of the world that is available. Freely drawing on this theory we do epistemology, i.e. we investigate the way in which human beings – including ourselves – come to formulate theories and to posit the existence of objects; among the theories thus investigated are those that we are drawing on in the course of our investigation. Now the crucial point is that to call a body of knowledge a "theory," or to call an object a "posit," does not in the least impugn its truth or its reality. "Theory" here is not mere theory, contrasted with real knowledge, for *any* body of knowledge will count as a theory from the point of view of epistemology; nor is "posited object" contrasted with real object. Thus, as Quine famously says:

> To call a posit a posit is not to patronize it . . . Everything to which we concede existence is a posit from the standpoint of a description of the theory-building process, and simultaneously real from the standpoint of the theory that is being built. *Nor let us look down on the standpoint of the theory as make-believe; for we can never do better than occupy the standpoint of some theory or other, the best we can muster at the time.* (1960: 22; my emphasis).

The crucial point is the one emphasized: there is no alternative to occupying some substantive theory of the world, and to do this means accepting that theory, at least for the moment, as true, and accepting its objects as real. Of course we may develop our theory into a different one, but we cannot occupy some neutral philosophical vantage point; nor can we accept a theory while still pretending that we are not accepting it as true.

In Quine's use, then, theoretical knowledge is not contrasted with ordinary knowledge. Similarly, theory for Quine is not contrasted with fact. All knowledge is theoretical, in Quine's sense. Just as there is here a stretching (or a distortion) of the word "theory," so also there is a stretching of the word "posit." When we say of, neutrinos, for example, that they are posits, we would generally be taken to mean that some person or group of people consciously posited them. Quine, however, speaks of physical objects in general as posits, and here there is no such implication. No conscious decision was ever taken to posit such things. As in the case of "theory," Quine's use thus assimilates ideas which one might suppose to be importantly different.

Metaphysics and regimentation: logic and extensionality

To this point we have seen Quine reflecting on the nature of our knowledge, its sources and bases, and on its status. The philosopher's task, as he conceives it, also includes clarifying our knowledge and helping us to attain a clear view of just what it comes to and what it really commits us to. This latter kind of task may be thought of as the Quinean version or analogue of metaphysics. Like the others, it aims, in Quine's view, to contribute to the overall scientific enterprise. It should, for example, enable us to avoid useless or misleading questions, help to suggest fruitful lines of further inquiry, and expose potential problems that may lurk in scientific theories.

How is this task to be approached? The method here is to show how various parts of our knowledge could be reformulated in the clearest possible terms. We have reason to take the objects and categories revealed by this reformulation as real and fundamental. An extended passage from *Word and Object* is worth quoting at length on this topic:

> The same motives that impel scientists to seek ever simpler and clearer theories adequate to the subject matter of their special sciences are motives for simplification and clarification of the broader framework shared by all the sciences. Here the objective is called philosophical, because of the breadth of the framework concerned; but the motivation is the same. The quest of a simplest, clearest overall pattern of canonical notation is not to be distinguished from a quest of ultimate categories, a limning of the most general traits of reality. Nor let it be retorted that such constructions are conventional affairs not dictated by reality; for may not the same be said of a physical theory. True, such is the nature of reality that one physical theory will get us around better than another; but similarly for canonical notation. (1960: 161)

Themes that we examined in the first part of this essay, having to do especially with the rejection of the Principle of Tolerance and with the remoteness of some parts of our theory from experience, re-emerge here. Simplicity, clarity, and convenience are not merely "pragmatic" virtues which are to be distinguished from the cognitive or theoretical virtue of truth. If we have a language which displays those virtues to the highest extent, then we have reason to take the structure of that language as telling us something about the real world, just as we take theories which display those virtues to tell us something about the world.

One aspect of the philosopher's task, then, is to find the clearest and simplest framework in which to formulate our knowledge. In Quine's view, this clarity and simplicity is to be found in first-order logic. (By "first-order" logic is meant logic of truth-functions and quantifiers which bind variables in positions occupied by singular terms, but not positions occupied by predicates.) He takes it that our knowledge is at its clearest when it is formulated in the syntax of first-order logic: a syntax which uses only truth-functions, predicates, variables, and quantifiers. This is not to say that he advocates language reform, or that he thinks that we should in fact reformulate all of our knowledge in those terms. But where our concern is with getting clear about what some part of our knowledge really commits us to, we would do well to consider how it might be phrased in logical syntax. This syntax is extraordinarily transparent and economical, which makes theorizing about it simple, and yet has surprising expressive power. It contains neither proper names nor function-symbols among its primitive expressions, for example, yet the effect of each can be easily achieved. (The basic techniques of doing so derive from Russell's Theory of Descriptions (see RUSSELL).) One moral to be drawn here is that we may be able to achieve the effect of a particular construction without in fact having to augment our stock of primitives. Quine is much concerned to take advantage of this kind of economy wherever it is possible, for the sake of the clarity and simplicity of the overall theory. (In ontology, as we shall see, Quine has a similar concern with economy: to show how particular kinds of object which appear to be assumed in what we take ourselves to know need not in fact be taken for granted as primitive, because their effects can be duplicated by other means.)

One feature of this framework which has proved extremely controversial is its *extensionality*. A language is extensional when any sentence of it retains its truth-value under any one (or more) of three kinds of changes: (1) any name in it may be replaced by any other name of the same object; (2) any predicate in it may be replaced by a co-extensive predicate (i.e. one true and false of exactly the same objects); (3) any sentence embedded in another sentence may be replaced by a sentence of the same truth-value. It is important to note that these three requirements which a language must satisfy to count as extensional interlock in ways that make it very hard for a reasonably comprehensive language to satisfy any one of them without satisfying all three. This point is important because the first requirement is, on the face of it, far more plausible than the third. It is extremely plausible that if we really understand what a given sentence is about, then we could replace the name of that object in the sentence with another name of the same object without altering the truth-value of the original. If a sentence is genuinely *about* an object, how we name that object should be a matter of indifference to the truth or falsehood of the sentence, though it may affect, for example, its poetic quality. (One might, indeed, take this idea as a partial definition of the somewhat

vague idea of *aboutness*.) No such superficial plausibility attaches to the third requirement, yet it turns out to be very hard to see how one can accept the first without also accepting the third. (See, for example, Quine 1995: 91–2.)

Any language which uses the syntax of first-order logic (along with any standard semantics) is, in virtue of that fact, an extensional language. Yet it would be a mistake to think Quine accepts extensionality simply in order to be able to use the syntax of logic. On the contrary: he takes one of the advantages of that syntax to be that it enforces extensionality, which he holds to be desirable for its own sake:

> I find extensionality necessary . . . though not sufficient, for my full understanding of a theory. In particular, it is an affront to common sense to see a true sentence go false when a singular term in it is supplanted by another that names the same thing. What is true of a thing is true of it, surely, under any name. (1995: 90–1)

Quine's insistence on extensionality is a very long-running theme of his thought, going back as far as the clarification of *Principia Mathematica* (Whitehead and Russell 1910–13) in his doctoral dissertation. (See the first paragraph of this essay; also Quine 1991: 265–6.)

The requirement of extensionality has been controversial because there are large areas of discourse which, at least if taken at face value, are not extensional. This for Quine is reason enough not to take such discourse at face value. Instead, he thinks, we should either try to reformulate it so that it becomes extensional or else exclude it from the more scientific and respectable parts of our knowledge, those parts which we take as really telling us about the objective world. We shall discuss three examples; our discussion of each will be very brief, although the second and third of them are issues which have generated much controversy, and on which Quine has, largely for that reason, written extensively.

The first sort of prima-facie violation of extensionality is almost trivial; indeed it might be said that the idea that we really have such a violation here is simply a mistake. This is the case of quotation, which is worth examining because it functions as something of a paradigm for Quine. It is true to say: " 'Quine' has one syllable"; it is false to say " 'The author of *Word and Object*' has one syllable"; yet the one sentence might seem to be obtained from the other by replacing a singular term with another singular term designating the same object, since Quine, of course, *is* the author of *Word and Object*. Here the solution to the apparent puzzle is easy. We should not construe the subjects of the sentences as referring to a person, but rather to the words, the expressions themselves. It is of the *word* "Quine," not of the philosopher Quine, that we say it has one syllable. And then of course the substitution no longer replaces a name by another name for the same object, since the two *expressions* are not the same.

The second sort of example is far less easy to dismiss. The case of indirect discourse, where one reports the speech or the thoughts of another, also gives rise to prima-facie cases of non-extensionality. Othello (supposing him for the moment to be a real person) has the false belief that Desdemona loves Cassio, but does not (presumably) have the equally false belief that the moon is made of green cheese. Clearly, in statements of the form "*A* believes that *p*" it is not only the truth-value of *p* that is relevant to the truth-value of the whole. Similarly in the case of singular terms. To use an example of

Quine's, Tom may believe that Cicero denounced Catiline without believing that Tully denounced Catiline, for he may not know that Cicero is Tully.

One response to such cases is to say that what is believed or disbelieved is not a fact about a person (or other object) but rather a *proposition*. (For this reason, philosophers often speak of such cases as statements of "propositional attitudes.") Then it is claimed that we have two distinct propositions: first, that Cicero denounced Catiline, and, second, that Tully denounced Catiline. So construed, belief-contexts become extensional, for we can no longer obtain a falsehood from a truth by substituting a co-designative expression (since this now means an expression referring to the same proposition). Quine rejects this idea, chiefly on the grounds that no clear and precise identity-conditions have been given for propositions, which should therefore not be accepted for scientific and philosophical purposes. This is of a piece with his having some degree of skepticism about meaning, for one can think of a proposition as being the meaning of a declarative sentence. (In that case the question of the identity-conditions of propositions is the same as the issue of synonymy for declarative sentences.)

Quine's treatment of indirect discourse relies not on the meanings of sentences but simply on the sentences themselves. Formally, he assimilates them to cases of quotation, by construing belief (and doubt, and hope, and so on) as attitudes towards *sentences*. Perhaps there is something odd or counterintuitive about speaking of believing or not believing a sentence, but this kind of oddity is something that Quine is, as we have seen, fully prepared to accept in the interests of clarity (as he conceives it). And if we do talk this way, then there is nothing puzzling in thinking that Tom, to revert to our example, may believe the one sentence without believing the other, for they are distinct objects.

Two points call for comment. First, Quine was for a period convinced that there are two kinds of belief. There is the ordinary kind, variously construed as an attitude towards a sentence or a proposition. But then there is also, he thought, another kind, *de re* belief, which is belief genuinely about an object, so that the way in which the object is described is irrelevant to the truth of the belief-attribution. He spent much effort trying to make sense of cases which appear to be of this sort, but subsequently came to think (quite correctly, in my view) that the appearance of two distinct kinds of belief is mistaken, and that the phenomena that he was attempting to understand are in fact not really cases of a different kind of belief. Second, it would be a mistake to think of the difficulties of making clear sense of belief contexts as merely formal. It is Quine's view, after all, that we want an extensional language not for its own sake alone, but because lack of extensionality is a sign of lack of clarity. His re-construal of belief as an attitude towards sentences puts the emphasis in what he thinks, for independent reasons, in the right place. If we are construing evidence as austerely and strictly as possible, the evidence for a belief-ascription is simply a report of what the person concerned said – the very words uttered – and did. In reports of these sorts there is no violation of extensionality. Beyond that, belief-ascription relies on empathetic projection: the ascriber imagines what it would be like to be in the believer's situation. Here, it may be thought, we are out beyond the realm of hard fact.

A third area of discourse in which extensionality fails, at least on Quine's view, is that concerning modality. Suppose it said that nine is *necessarily* greater than five (the example, again, is Quine's). Replacing "nine" by a co-designative expression we can

obtain: the number of the planets is necessarily greater than five. Quine, however, sees no hope of making sense of any notion of necessity according to which this is true (cf. MARCUS). He distinguishes a notion of necessity of this sort from one which attaches to sentences (and hence can be assimilated to the case of quotation) or treats it as a statement operator. In these cases the necessity is supposed to hold of the sentence, not of the object, so there is no risk of the sort of violation of extensionality indicated above. The general issue of modality, however, is not a crucial one from a Quinean point of view, for he sees no need to accept any notions of necessity or possibility. He holds that a reconstruction of our knowledge, or of those aspects of it which we want to take with full seriousness, as telling us about the objective world, require no such notions.

We have briefly discussed three cases of prima-facie non-extensionality. It is worth emphasizing, however, that these three do not exhaust the matter. Counterfactual conditionals, statements of causality, and, if taken at face value, statements of time and tense, are among other such cases.

Ontology and its relativity

A crucial part of the task of getting clear about what our theories commit us to, on Quine's account, is gauging their *ontological commitments*. By the ontological commitments of a theory Quine means what entities that theory says there are in the world. In Quine's view, the way to settle this is by seeing what objects must be in the range of the theory's variables if it is to be true. Quine's emphasis on ontology, and on the range of variables as the measure of ontological commitment, is perhaps in part to be explained by his early work in set-theory, and his abiding interest in that subject. In that context, the range of the variables of the theory is a natural measure of its strength, and the threat of paradox makes this matter of vital concern. (The importance of this point was emphasized to me by Stephen Menn, to whom I am grateful.)

Quine takes ontology to be a product of self-conscious scientific and philosophical reflection. He does not see the philosopher's ontological task as that of capturing in perspicuous form the ontology implicit in ordinary thought and discourse, for he insists that there is no such "ordinary ontology":

> a fenced ontology is just not implicit in ordinary language. The idea of a boundary between being and non-being is a philosophical idea, an idea of technical science in a broad sense. Scientists and philosophers seek a comprehensive system of the world, and one that is oriented to reference even more squarely and utterly than ordinary language. Ontological concern is not a correction of a lay thought and practice; it is foreign to the lay culture, though an outgrowth of it. (1981: 9)

When we are concerned with ontological questions, or metaphysical questions more generally, we cannot simply examine our beliefs in the terms in which we are at first prone to express them. Our beliefs must, rather, be cast into a standard notation, which will let their presuppositions shine forth. Ontology, as Quine interprets it, thus presupposes regimentation: it is only insofar as we conceive of our knowledge as cast in regimented notation that it makes sense to raise ontological questions.

The artificiality of ontology, in Quine's view, is an important point. He is often criticized for distorting ordinary thought, or doing violence to our supposed "intuitions." On his view these criticisms miss their mark entirely. More positively, he takes part of the task of ontology to be that of showing just what our theories really commit us to. Here he has an interest in ontological economy. In many cases our theories seem to commit us to accepting entities of a certain kind, but artful re-construal of the theories shows that in fact we need not presuppose entities of the given sort. Here artificiality is inevitable.

In this spirit, Quine takes the reduction of numbers to sets, and of ordered pairs to sets, to be clear philosophical achievements. (Section 53 of Quine 1960 is entitled "The Ordered Pair as Philosophical Paradigm.") In each case the reduction shows how our discourse could be rephrased so as to avoid commitment to a kind of entity to which it appears to be committed. The rephrasing is not meant as a practical substitute for normal arithmetical or set-theoretic language. Nor is it claimed that the paraphrased version gets at the "real meaning" or hidden structure of the original. It is simply that the reduction shows us how we could do everything that we need, for scientific purposes, without presupposing that there are ordered pairs (or numbers), and hence that we need not take ourselves to be committed to the existence of such things. Nor is it only towards such technical subjects as arithmetic and set-theory that Quine takes this attitude. On the contrary; he takes the same view everywhere. He holds, for example, that we can eliminate minds in terms of bodies, just as we can eliminate numbers in terms of sets. Instead of speaking of a person's mind at a given moment, we could instead speak of his or her body at that moment. Mentalistic predicates, such as "is thinking of Vienna" persist, but no exclusively mental *objects*. The maneuver is trivial, but not, in Quine's view, any the worse for that; he takes it as showing that we have no ontological commitment to minds, over and above bodies.

Quine thus stresses the artificiality of ontology, and the use of paraphrase or re-construal to eliminate apparent ontological commitments. His tactics here presuppose something implicit in our earlier discussion of knowledge, especially of observation sentences. In his view language is, of course, referential – it is *about* things. We refer to people and other objects in almost everything we say. But the relation of reference is not our fundamental cognitive relation to the world. The fundamental relation, the way that language gets to be about the world at all, is the relation of observation sentences to patterns of stimulation. (In particular, the relation of a given observation sentence to the patterns of stimulation on the basis of which speakers of the language are inclined to accept or reject the given sentence.) This is not a relation of reference, for observation sentences are not *about* patterns of stimulation. Reference is a derivative relation. Indeed the very notion of an object, that to which we refer, is in Quine's view derivative – though it comes so naturally to us that it seems inevitable. If there were a language consisting only of observation sentences, there would be no reason to attribute any reference to that language. The vastly greater complexity of our language requires inferential links between sentences, links which we can grasp only by attributing structure to the sentences. Seeing sentences as divided into terms, some of which refer, is part of this process:

> Reference and ontology recede thus to the status of mere auxiliaries. True sentences, observational and theoretical, are the alpha and the omega of the scientific enterprise. They are related by structure, and objects figure as mere nodes of the structure. (Quine 1990: 31)

The derivative character of ontology manifests itself most dramatically in the Quinean doctrine known as *ontological relativity*, or the *inscrutability of reference*. Quine's claim is that it would be possible to carry out a large-scale re-construal of our knowledge, replacing every referring term by another. We could, for example, replace every term referring to a physical object by a term referring to its space-time complement, that is, to the whole of space and time *other* than the given object. Along with that we would re-construe our predicates, so that a predicate true of a given object would now be taken as true of its space-time complement. These re-construals would cancel out, leaving the truth-value of each sentence the same as before. The result would be a system of knowledge exactly like our own in its structure, including its relation to evidence, but in which each term refers to objects different from those to which its unreconstructed analogue refers. Nothing, Quine claims, prevents such a re-construal. The passage quoted in the previous paragraph continues: "What particular objects there may be is indifferent to the truth of observation sentences, indifferent to the support they lend to theoretical sentences, indifferent to the success of the theory in its predictions" (Quine 1990: 31).

Quine claims that this sort of wholesale re-construal of our knowledge is possible, that it would not be inconsistent with any part of it. True, I can and would stoutly maintain that my word "rabbit," say, refers to rabbits and not to their space-time complements. But *all* my uses of the word, including those uses in which I insist that it is rabbits that I am talking about, are open to re-construal. From the possibility of this sort of re-construal, Quine sometimes infers that there is no fact of the matter about reference – no fact of the matter as to whether my word "rabbit," say *really* refers to rabbits, or rather to the space-time complement of all the rabbits. The thesis of ontological relativity, however, is perhaps best thought of as a rejection of this idea of "really refers," insofar as it outruns the ordinary idea of reference. Taking our own language for granted we can say to what the words of another language (or, trivially, of that same language) refer; while we stay within that language we use it to refer with no more than the usual difficulties or ambiguities. Ontological relativity is, Quine says, "unproblematic but trivial" (Hahn and Schilpp 1998: 460). Only when we translate, or map our language onto itself, does the ontological relativity emerge. And here it should perhaps be seen as a reminder of the derivative status of reference.

Quine has come to give increasing emphasis to the doctrine of ontological relativity. Should we see it as conflicting with, or undermining, his insistence on realism? Since it leaves truth-values unaffected, it does not affect his claims to be a realist about truth. That latter claim, as we saw, was also one that might be doubted. Quine's claim there is perhaps that he is as much of a realist as it makes sense to be: that there simply is no coherent sense of realism stronger than his. Similarly here, perhaps. Quine claims that ontological relativity is beyond doubt; it "admits of trivial proof," he says (Hahn and Schilpp 1998: 728). So there is in his view no chance of defending a version of realism that denies it. Ontological relativity might thus be taken as showing us what realism

can come to. Ontology simply is derivative upon truth, and given any system of truths there simply will be more than one way of construing its ontology. If this undermines our previous conception of realism then, from Quine's point of view, so much the worse for that conception.

Conclusion

I shall not attempt to summarize what is already a very compressed treatment of Quine's thought. It is perhaps worth saying, however, that I have ignored, or treated very briefly, a number of issues that have occupied considerable space, both in Quine's own writings and in the works of commentators and critics. Notable examples here include the indeterminacy of translation, Quine's views of indirect discourse and of modality, the question of his physicalism, and what it amounts to, and his substantive views on ontology. My decisions as to how to use the limited space available here are of course based on my view of what is most important in Quine's thought. But the reader should perhaps know that others might have made these decisions rather differently.[1]

Note

1 For their comments on earlier drafts of this essay I am indebted to Bill Hart, Peter Hacker, Gary Kemp, and Al Martinich.

Bibliography

Works by Quine

1936: "Truth by Convention," in *Philosophical Essays for A. N. Whitehead*, ed. O. H. Lee, New York: Longmans. (Reprinted in Quine, *The Ways of Paradox*, rev. edn., Cambridge, MA: Harvard University Press, 1976.)
1960: *Word and Object*, Cambridge, MA: MIT Press.
1961: *From a Logical Point of View*, 2nd, rev., edn., Cambridge, MA: Harvard University Press. (First published 1953.)
1963: "Carnap and Logical Truth," in Paul A. Schilpp (ed.) *The Philosophy of Rudolf Carnap*, La Salle, IL: Open Court. (Reprinted in Quine, *The Ways of Paradox*, 1976.)
1969: *Ontological Relativity and Other Essays*, New York: Columbia University Press.
1974: *Roots of Reference*, La Salle, IL: Open Court.
1981: *Theories and Things*, Cambridge, MA: Harvard University Press.
1987: *Quiddities: An Intermittently Philosophical Dictionary*, Cambridge, MA: Harvard University Press.
1990: *Pursuit of Truth*, Cambridge, MA: Harvard University Press.
1991: "Two Dogmas in Retrospect," *Canadian Journal of Philosophy* 21, pp. 265–74.
1995: *From Stimulus to Science*, Cambridge, MA: Harvard University Press.
2000: "Response to O'Grady," *The Proceedings of the World Congress of Philosophy, Boston, 1998*, Bowling Green, OH: Philosophy Documentation Center.

Works by other authors

Carnap, Rudolf (1934) *Die logische Syntax der Sprache*; Vienna: Springer Verlag. (Published, with additions, as *The Logical Syntax of Language*, trans. Amethe Smeaton, London: Routledge and Kegan Paul, 1937.)

Creath, Richard (ed.) (1990) *Dear Carnap, Dear Van*, Berkeley and Los Angeles: University of California Press.

Gibson, Roger (1982) *The Philosophy of W. V. Quine*, Tampa: University Presses of South Florida.

Gregory, Paul (1999) "Quine's Natural Epistemology," Ph.D. dissertation, University of Illinois.

Hahn, Edwin and Schilpp, P. A. (eds.) (1998) *The Philosophy of W. V. Quine*, expanded edition, Chicago and La Salle, IL: Open Court Press. (Previously published 1986.)

Hylton, Peter (1990–1) "Translation, Meaning, and Self-Knowledge," *Proceedings of the Aristotelian Society* 91.

Lee, Harold N. (1998) "Discourse and Event: The Logician and Reality," in Hahn and Schilpp, pp. 295–315.

Smart, J. J. C. (1969) "Quine's Philosophy of Science," in *Words and Objections: Essays on the Work of W. V. Quine*, ed. Donald Davidson and Jaakko Hintikka, Dordrecht: Reidel, pp. 3–13.

Whitehead, Alfred North and Russell, Bertrand (1910–13) *Principia Mathematica*, 3 vols, Cambridge: Cambridge University Press. 2nd edn. 1927.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

16

A. J. Ayer (1910–1989)

T. L. S. SPRIGGE

Language, Truth and Logic

General character of the book

A. J. Ayer rose to early philosophical fame with the publication in 1936, when he was 25 years old, of what remained his most famous, or infamous, book, *Language, Truth and Logic*. The work is his own version of the logical positivism characteristic of the Vienna Circle (whose meetings he had attended for three months in 1932–3), his outlook being closest to that of their leader, Moritz Schlick. The book is also strongly influenced by the British empiricist tradition, in particular by Hume and Russell. It was something of a bombshell to British philosophers and became for them the paradigm statement of logical positivism, threatening the outlook of some, providing an exciting intellectual liberation for others.

The book opens with the striking statement:

> The traditional disputes of philosophers are, for the most part, as unwarranted as they are unfruitful. The surest way to end them is to establish beyond question what should be the purpose and method of philosophical enquiry. (1946: 33)

So far as philosophy goes, Ayer's concern is to show the meaninglessness of metaphysical theories about a reality beyond the empirical. More generally, he also claims to show that religious statements, as usually now intended, are meaningless, as also are statements of fundamental ethical principle (except as mere expressions of emotion).[1]

To establish the meaninglessness of all such statements Ayer puts forward the verification principle. According to this a statement

> is factually significant to any given person, if, and only if, he knows how to verify the proposition which it purports to express – that is, if he knows what observations would lead him, under certain conditions, to accept the proposition as being true, or reject it as being false. If, on the other hand, the putative proposition is of such a character that the assumption of its truth, or falsehood, is consistent with any assumption whatsoever concerning the nature of his future experience, then, as far as he is concerned,

> it is, if not a tautology, a mere pseudo-proposition. The sentence expressing it may be emotionally significant to him; but it is not literally significant. (1946: 35)

Thus a meaningful statement must either be empirically verifiable, or be a tautology, that is, analytic or true by definition.

The passage just quoted is supposed to be a "somewhat vague" formulation of a principle that Ayer proceeds to express more precisely. But actually, since it turned out difficult to find a satisfactory precise formulation, it remains as good a formulation as any.

Since there are two types of meaningful statement for Ayer, the empirically verifiable and the analytic, his account of each of these will be considered in turn.

Empirical statements

Ayer distinguishes between strong and weak verification. A strongly verifiable proposition is one which could be conclusively established by sense experience, a weakly verifiable proposition is one which could be made probable by sense experience. It is too much to demand strong verifiability of a meaningful factual statement (it is doubtful indeed if any proposition is strongly verifiable) and so some form of weak verifiability is the appropriate criterion. Ayer tries to give an exact formulation of this as follows.

> Let us call a proposition which records an actual or possible observation an experiential proposition. Then we may say that it is the mark of a genuine factual proposition, not that it should be equivalent to an experiential proposition, or any finite number of experiential propositions, but simply that some experiential proposition can be deduced from it in conjunction with certain other premises without being deducible from these other premises alone. (1946: 38–9)

The general idea, here, is clear enough. A meaningful empirical statement must be a genuine aid to the anticipation of the experiences we can expect to have under various circumstances (identified in terms of the other experiences then available), though it need not tell us what experiences to expect all on its own. To illustrate his point Ayer gives an example of two questions that might be raised about a painting. (1) Was it painted by Goya? (2) Is the painting a set of ideas in God's mind? People may disagree in their answers to each of these questions, but in the first case they know what kind of empirical evidence would support their claim against that of their opponents, in the latter they do not (1946: 40).

Later in the work, especially in chapter VIII, "Solutions of Outstanding Philosophical Disputes," Ayer shows, or claims to show, how metaphysical questions are all meaningless in much the same manner, unless they are understood, as is often appropriate, as misleading ways of discussing how propositions of a certain type are to be analyzed. (See the section below, "What is the task of philosophy?")

Unfortunately Ayer later discovered that the technical formulation of this revelation of the meaninglessness of metaphysical questions was unsatisfactory. This is because any statement whatever, call it "P," can meet the condition simply in virtue of the fact that its conjunction with "If P, then O" (where "O" is a experiential statement)[2] entails

"O," as "If P then O" does not alone. Thus "God is annoyed," which Ayer would hate to find meaningful, entails an observation statement "You will shortly hear thunder" when conjoined with "If God is annoyed with what you said, you will shortly hear thunder." In his 1946 introduction to the second edition Ayer offered a more complicated formulation, which, however, he had later to concede, fell foul of technical criticisms from Alonzo Church and C. G. Hempel. (See Church 1949 and Hempel 1959.)

If one is not too infatuated with semi-formalization, however, one can, surely, say clearly enough what Ayer was getting at, whether one accepts it or not. Surely the real point of the verification principle, so far as factual (non-analytic) statements go, was this. Such a statement is meaningful to a particular individual if and only if it is possible for either it or its negation to be a practical aid to him in forming correct expectations about what he is liable to experience in the future. If there is no such possibility then it is factually meaningless, however much he may suppose himself to understand it.[3]

If there is a problem, here, it is about what "possible" means, but perhaps it is sufficient that the individual does not utterly rule out its occurring. It is to be noted, however, in this connection, that Ayer is anxious to distinguish practical verifiability from verifiability in principle . Thus, in an intriguingly dated example, "There are mountains on the other side of the moon" was said to be unverifiable in practice but verifiable in principle (an example taken from Moritz Schlick).

An important question is whether the verification principle is intended not only to tell us whether a statement is meaningful or not, but also to tell us what its meaning is. In effect, rather than in actual formulation, Ayer treats it as doing so and surely this is correct. For if a factually meaningful statement must be a possible aid to knowing what experiences to expect under various circumstances then its meaning must lie in the totality of such aid as it is capable of giving. If there is some residue of purported further meaning it would seem that this could be creamed off as an unverifiable statement included within it.

That the verification principle is intended to exhibit the meaning of factual propositions is plain from Ayer's deductions from it concerning the analysis of a whole range of ordinary statements of fact. Thus the verification principle is said to make inevitable a phenomenalist analysis of statements about material objects, since it is only "by the occurrence of certain sense-contents that the existence of any material thing can ever be in the least verified" (1946: 53). It is no good some objector saying that the existence of a physical object is not merely a fact about what sense-contents are available to us, though it is by this that it is verified, for ultimately all that can be verified by facts about sense-contents are facts about sense-contents and the probable truth of what can be inferred from such facts inductively. (See the following section). Ayer, however, like Quine later, eschews talk of meanings as entities, substituting for talk of meanings talk of synonymy (1946: 68) (see QUINE).

Analytic or a priori statements

The other sort of meaningful statements for Ayer were analytic statements. All genuinely necessary or a priori statements are of this type; thus anything like the synthetic a priori of Kant and others is rejected.

One initial point worth remarking is that, while Ayer is clear that, if a statement is empirically meaningful then so are its contraries and contradictory, his assertion that the only two types of meaningful statements are empirical hypotheses and analytic propositions, taken strictly, implies that this is not so in the case of the latter. If so, while "5 + 3 = 8" is meaningful, the proposition "5 + 3 = 9" is not false, but meaningless. How far this is intended is unclear, since Ayer does, in fact, talk of false mathematical statements. (See 1946: 86.)

Ayer's discussion of analytic propositions (in 1946: ch. IV) starts out from the problem which a priori truth is supposed to pose for empiricism (of which his logical positivism is avowedly a species). For empiricism can countenance no claim to knowledge that is not based upon sense experience, and even then what is called "knowledge" is always probable hypothesis rather than absolute certainty.

> Where the empiricist does encounter difficulty is in connection with the truths of formal logic and mathematics. For whereas a scientific generalisation is readily admitted to be fallible, the truths of mathematics and logic appear to everyone to be necessary and certain. But if empiricism is correct no proposition which has a factual content can be necessary or certain. Accordingly the empiricist must deal with the truths of logic and mathematics in one of the two following ways: he must say either that they are not necessary truths, in which case he must account for the universal conviction that they are; or he must say that they have no factual content, and then he must explain how a proposition which is empty of all factual content can be true and useful and surprising. (1946: 72–3)

Having dismissed the first alternative, that of J. S. Mill, according to which, for example $2 \times 3 = 6$ is simply so well confirmed a statement of fact that we (wrongly) think that it could not have been otherwise, Ayer opts for the view that all a priori and necessary (these are identified) so-called truths are really analytic.

Rejecting Kant's account of analyticity for various reasons, Ayer formulates his own account. This, however, is somewhat shifting. The most definitive formulation would seem to be this:

> a proposition is analytic when its validity depends solely on the definitions of the symbols it contains, and synthetic when its validity is determined by the facts of experience. Thus, the proposition "There are ants which have established a system of slavery" is a synthetic proposition. For we cannot tell whether it is true or false merely by considering the definitions of the symbols which constitute it. We have to resort to actual observation of the behaviour of ants. On the other hand, the proposition "Either some ants are parasitic or none are" is an analytic proposition. For one need not resort to observation to discover that there either are or are not ants which are parasitic. If one knows what is the function of the words "either," "or," and "not," then one can see that any proposition of the form "Either *p* is true or *p* is not true" is valid, independently of experience. Accordingly, all such propositions are analytic. . . . However, when . . . we say that analytic propositions are devoid of factual content, and consequently that they say nothing, we are not suggesting that they are senseless in the way that metaphysical utterances are senseless. For although they give us no information about any empirical situation, they do enlighten us by illustrating the way in which we use symbols. (1946: 78–9)

There are several difficulties with propositions of the type he classifies as both "analytic" and "a priori." On his account, are these propositions not, in effect, statements about how certain symbols are normally used? But if so, they seem to be empirical, since it is an empirical fact that we use words as we do. Ayer tackles this question in his introduction to the second edition. (See 1946: 16–18.) His reply is that, although they are simply the consequences of sticking to a certain consistent use of certain symbols, they do not so much state as presuppose such rules of language. And this, thinks Ayer, explains why they can be surprising. For there are (doubtless infinitely) many consequences of this sort which it requires considerable intellectual power to grasp.

What is troubling about this answer is that while the simpler statements that Ayer calls "analytic" may be thought of as little more than reminders of how we optionally use certain symbols, there are innumerable consequences of such use which follow therefrom in a manner that is not similarly optional. Compare the rules of chess. There is no proper answer, other than a presently irrelevant historical one, as to why the pieces may be moved just as they may, but from this set of optional rules untold consequences follow as to how the game can best be played to win. Or to take a case more to the point, even if the proposition that 7 is 6 + 1 is simply a reminder of the meaning of "7," the proposition that 7 is a prime number is not a reminder of how we use the symbol but a necessary consequence thereof. It is of this latter sort of necessity that Ayer gives no satisfactory account. Or so at least it seems to some of us.[4]

Be that as it may, the general idea is clear, namely that all necessary or a priori truths are really the consequences of an optional use of language and tell us nothing about anything non-linguistic. Nor do they exactly say anything about language, they simply helpfully reflect back to us the character of the language in which our knowledge or beliefs are expressed.

Perhaps Ayer's position is, in effect, that the function of analytic statements is not strictly to say anything but to serve as a kind of verbal drill whereby we reinforce and improve our command of the rules and accepted transformations which give verbal expressions their meaning.[5]

What is truth?

So much for language and logic. What of truth? One might suppose that for Ayer truth would consist in being a reliable predictor of sense experience in the case of empirical propositions, and following from linguistic rules in the case of analytic propositions. Actually, in chapter V Ayer puts forward what has been called the redundancy theory of truth. According to this, the correct account of what "true" means is that it simply emphasizes the assertion of the proposition said to be true. Thus to say that the proposition *that dogs bark* is true is simply to say that dogs bark. Similarly to say that a proposition is false is simply to assert its negation. To say that it is false *that cats bark* is simply to say that cats don't bark.

But how are we to explain occurrences of "true" where the proposition said to be true is not formulated, for example, "Everything he said in the lecture is true"? Ayer's implied answer is that this means "(p)(he said that p implies p)." There is a problem with this answer, into which we shall not enter, in that the variable "p" occurs firstly as a

name variable and secondly as a propositional variable; that is, the first occurrence of "*p*" stands in for the name of a proposition, and its second occurrence for the actual formulation of a proposition.

In the light of this Ayer says that philosophical attempts to answer the question "What is truth?" are largely misconceived. When they have a definite meaning, the question asked is really, "How are propositions *validated*?" The meaning of "validated" is not made clear, but the question seems to mean "What are the criteria we properly use in deciding whether to affirm or deny them?" (Ayer often uses "validate," as it seems to me, to avoid talk of *judging to be true*.) The answer is that analytic propositions are validated by their being consequences of the way we use words, while empirical propositions are validated by the fact that they have been found successful as a way of predicting what we will experience under various circumstances, and are thereby taken as likely to be similarly successful in the future. (See 1946: 99.) If we raise "the problem of induction" associated especially with Hume, as to what right we have to take past experience as a guide to the future, Ayer's answer is, roughly, that to be guided in this way is just what we presently mean by being "rational."

What is the task of philosophy?

If metaphysics is nonsense, and to be abandoned, is there any type of philosophy that is more intellectually respectable? Ayer's positive answer is that the sort of philosophy that is a worthwhile activity is (conceptual) analysis. (See 1946: chs II and III.) And in fact this was what genuinely great philosophers have always been mainly engaged in. Often they have put their questions in the form of "What is *X*?" e.g. "What is matter?", "What is time?", "What is the self?", and there is nothing wrong in this mode of expression, if it is properly understood. Thus understood, these questions are really requests for definitions of some of the very general expressions in our language which either puzzle us or lead to metaphysical nonsense.

The traditional type of definition professes to explicate the meaning of a word by offering some other more complex verbal expression which is its equivalent. Ayer's trivial example is "An oculist is an eye doctor," which tells us that "oculist" and "eye doctor" mean the same (1946: 60). Definitions of this type are for the most part of limited use to philosophy, which is concerned with a more fundamental clarification of what both expressions are meant to stand for. Instead the philosopher requires so-called "definitions in use." (See pp. 60–3.) A definition of this sort is an instruction for translating statements about *X* or *X*s into equivalent statements which have no word or expression referring (grammatically considered) to *X* or *X*s. These are useful when *X*s strike a philosopher as somehow not belonging to the bedrock of reality.

Thus a philosopher might try to answer the question "What is a nation?" by showing how specimen statements about nations can be translated into statements about people. For example, he might seek a way of translating "Britain and Germany were at war from 1939 to 1945" into a complicated statement about how people whose homes were on one part of the earth's surface behaved towards, and were affected by, people whose homes were on another part of the earth's surface. If such a translation of statements about nations into statements about people and land is possible, then a nation may be described as a "logical construction" out of people and land, though it should be

realized that this is not a statement about how two sorts of *thing* are related but between *linguistic expressions*, one of which is supposedly puzzling in a way which the other is not.

Even in such rather obvious cases as that of nations such definitions are usually gestured towards rather than actually formulated. The gesture may be sufficient, however, to show that what the statement about the nations says in a simple, but *misleading* (because it suggests that nations are something over and above persons and land) way, is something which in principle should be sayable in a way that gives us no excuse for being thus misled. As for the difficulty in finding quite satisfactory actual translations, this, it may be suggested, may be because they would have to be impossibly complicated, or because they would have to be precise about details that the statements being explained leave vague. So although war is certainly a matter of people doing things one cannot be precise about just what people must do to be at war. (Cf. 1954: 141–3.)

So the task of philosophy is to point towards definitions in use of expressions that are liable to puzzle us or to suggest that there are things over and above those we actually encounter empirically.

Phenomenalism regarding physical objects

Central to Ayer's type of logical positivism is a phenomenalist view of physical or material objects. According to this, every statement about the physical world is, in principle, translatable into a proposition to the effect that under such and such conditions such and such sensations will or would occur. Thus the proposition that a physical thing exists always means "that, if certain conditions were fulfilled, certain sense-contents . . . would be experienced" (1946: 141).[6]

"Experienced by whom?" one may well ask. The answer would seem to be "by whomsoever it is who is affirming the proposition"; in short, when you affirm it, it tells you what sensations you should expect under such and such circumstances, while when I affirm it, it tells me what sensations I should expect under such and such circumstances. This suggests that each of us gives our own private meaning to the proposition. And in fact the doctrine of the book, without perhaps the author being fully aware of it, is that what counts as the same factual statement has a different meaning for each person, since for each of them the information it provides, if it is true, concerns just their own actual and possible experience. This is disguised, somewhat, by the fact that persons are themselves supposed to be logical constructions out of sensations, though not the same sort of logical construction as physical objects are.

For a person, so far as his conscious mind goes, is, according to Ayer, a logical construction out of those sense-contents that occur in the same sense-fields as do the organic sense impressions of that body. Among the consequences that follow from this is that an individual's survival of bodily death is a meaningless idea, insofar as there can be no organic sensations of his body thereafter. Here Ayer differed from the leader of the Vienna Circle, Moritz Schlick, who thought it perfectly meaningful to suppose that I might verify my own death by having sense impressions as of seeing my funeral from a point of view unoccupied by a human body. (See Schlick 1949: 159–60.)

Propositions about other minds

This leads naturally to the account of our knowledge of other minds presented in *Language, Truth and Logic*, chapter VII. It is likely to be charged, he says in effect, that if I accept the doctrine of that book then I am committed to solipsism, to the view that only I exist, as a conscious individual with my own sense experiences. For the existence of the sense-experiences of other people appears to be, in principle, something that I cannot verify, since my sense-impressions can only be associated with the organic sensations of my own body, not with those of another person. But Ayer rejects this conclusion, contending that each of us must define the existence of other persons, including the sensations that go with the organic sensations of their body, in terms of the behavior on the basis of which I would ordinarily conclude that they were conscious and had sensations.

Thus he held, at that stage, that what "I" (whoever I am) mean when I speak of my own sensations is that such and such sensations actually occur together with the organic sensations of my body, but that when I speak of the sensations of another person I mean that they are behaving, or are disposed to behave, in such and such a way (this being a physical fact ultimately consisting in facts about my sensations as of perceiving their bodies move and make noises etc. of such and such a sort).

This view about what we mean by speaking about the sensations of others, reveals, as Ayer later himself insisted, a peculiar double-take on the whole business of what we mean by what we say (see 1956: 245–7). The official view of *Language, Truth and Logic* is that everyone means by everything they say something about their own sensations. But this is a view delivered as true of every speaker, by Ayer, *qua* philosopher, who, thereby, is clearly supposing that other people have sensations in the same sense as he does (i.e. in a way not analyzable behavioristically) though it is part of the theory that this realist conception of the sensations of others is, for each of us (and that must include Ayer himself) meaningless (see 1946: 141).

Propositions about the past

As strange, or stranger, than this view of the meaning of assertions about the experiences of others is the view that propositions about the past can only be meaningful (because otherwise unverifiable in principle) if they are equivalent to predictions about the kind of so-called historical evidence that would support them. An oddity (we may remark) of this view is that, while empirical knowledge is said to consist in predictions about future experience on the basis of past experience, the fact that the past experiences occurred is itself a prediction of the same essential kind.

The paradoxical character of these conclusions is among the factors which led Ayer away from the precise positions of *Language, Truth and Logic*, though he struggled to remain true to at least the general spirit of the verification principle.

Critique of ethics and theology

One position, however, which Ayer never abandoned was the emotive theory of ethics advanced in chapter VI (see STEVENSON). According to this ethical concepts are pseudo-

concepts, that is, they lack factual meaning. To say that behavior of a certain sort is wrong is not to state anything about it which can be true or false. It is simply to express a feeling. True, there may be ethical statements which include a factual element. If I say "He did wrong to kill the cat," then, inasmuch as this says that he killed the cat it is meaningful, and true or false. But calling it wrong adds nothing cognitively meaningful. And if I make a statement of pure ethical principle, such as "Suicide is wrong" then this is just as though I held up my hand in horror at the idea of people committing suicide. This has been called (not by Ayer) the "boo-hurrah" theory of ethics. He gave it a milder statement in a later essay (1954: essay 10) pointing out that he was not making a negative value judgment about moral thinking, and, in fact, Ayer was personally and publicly committed to strong liberal principles.

As for religious propositions, statements like "God exists" or "God loves us," as most people now think that they understand them, they are meaningless. There is no observational test that could be used to determine their truth or falsehood. God is not, as the mountains on the other side of the moon were then, something whose existence is in fact unverifiable, but in principle verifiable. It is worth noting that even if Ayer revised his view that life after death was meaningless he could still put up a good case for saying that the existence of God would remain so, for whatever experiences I might have in some other world, none of them (so Ayer could easily argue) would show that there was or was not a God as sophisticated monotheism describes Him. Of course, if God is conceived of as an enormously powerful being in human form, as depicted in religious paintings, the matter would be different, but a religious sophisticate will say that that is a mere image of a truth which cannot be expressed in sensory terms. At that point Ayer said that he was not an atheist, since "God does not exist" is as meaningless as "God exists." Later he relaxed this somewhat and was prepared to call himself an atheist on the grounds that no meaning can be given to the proposition that God exists that makes it remotely likely to be true. (See Ayer 1973.)

The future of philosophy

Officially in *Language, Truth and Logic* Ayer regarded the positive task left for philosophy – after the elimination of metaphysics and the final analysis of the statements of everyday life, which a verificationist approach like his had, at least *almost*, finally achieved (see pp. 152–3) – as the analysis of the concepts of science. But this was hardly undertaken in that work, and was never a main concern of Ayer's (except to some extent in his treatment of probability). His interest was always in those traditional questions of philosophy his radical answers to which we have been discussing, and most of his later work consists in attempts to find more persuasive answers to them in the light of a less extreme form of verificationism.

Later positions

It is unfair to take *Language, Truth and Logic* as being Ayer's main contribution to philosophy. There is much insufficiently admired later work. At present Ayer is out of fashion and undervalued. He would hardly have ground for complaint at this, scoffer as he was as a young man at his predecessors, but greater justice will be done to him one day.

He is probably right that his best book was *The Problem of Knowledge* (1956). This presents philosophical epistemology (theory of knowledge) as primarily concerned to explain what knowledge is by an examination of skeptical difficulties found in various ordinary claims to possess it. In the process he presents a revised view of such matters as we have considered in connection with *Language, Truth and Logic*, revisions to a great extent already presented in the essays collected in *Philosophical Essays* (1954).

Ayer contends that for someone to know that something is the case, it is required (1) that he feels sure that it is so, (2) that it is so, and (3) that he has the right to be sure that it is so. (See 1956: 34).

Skeptics have raised all sorts of doubts about our right ever to be sure of the truth of any propositions concerning physical objects, other minds, and past events, whether these propositions simply assert that there really are such things or say something more specific about particular cases.

The pattern of skepticism is as follows. A contrast is drawn between the evidence on which such propositions are believed and what they claim to be the case. Thus our evidence for propositions concerning physical objects always consists in facts about our own sense-data, for propositions concerning other minds in facts about the behavior of other organisms, and for propositions concerning past events in apparent memories or records of them.

There are four types of philosophical riposte to such skepticism.

(1) *Naive realism*. This denies that our evidence for the problematic knowledge claims *is* indirect in the manner the skeptic alleges, rather do we have a direct or immediate experience of physical things or other minds or the past (whichever is in question).

(2) *Reductionism*. This analyses the problematic truths into truths about the things which feature in the evidence for them. Thus facts about physical objects are reduced to facts about sense-data and facts about other minds to the behavior of others. The interpretations of such propositions in *Language, Truth and Logic* are paradigm cases of reductionism.

(3) *The scientific approach* (or *causal inference theory*). The problematic propositions are inferred as the causes of the things which are our evidence for them.

None of these approaches is altogether successful. Statements about physical objects cannot be translated into statements about sense-data nor statements about other people's experiences into facts about their behavior. Naive realism simply ducks the problem. And causal realism is vulnerable to the objection that a proper causal inference must be to a reality which we could know about more directly, and thus be able to check that the causal relation between our evidence and what we take it as evidence for really holds.

(4) *Descriptive analysis*. The sting of the problem as the descriptive analyst sees it lies in the fact that we complain that we have not got a way of knowing about these things when it is logically impossible that we should have. For example, the complaint that other minds are closed to us loses its alarming quality when we realize that it is a necessary truth (analytic) that our belief in anything must rest *on our own experience*, so that it is not that we lack some power which it would

even make sense of someone having. Once this is realized, we can be content with a careful description of our normal way of forming beliefs about the minds of others, since this is often as good as it makes sense to wish for. Similarly for the other problems.

Ayer's own favored strategy in each case is the last, that of descriptive analysis. However, it must be said that in the case of the first problem (that of our knowledge of physical objects, always a special interest of his) it is doubtful how far Ayer really means to distance himself from reductionism. For various somewhat technical reasons there can be no actual *translation* of statements about physical objects into if–then statements about sense-data, but the suggestion lingers in the air that somehow there is nothing else for them to *tell us about*. (See Ayer 1956: 131–47 and 1954: essay 8.)

Ayer's approach to the problem of other minds is somewhat different. (See 1956: 243–54.) The obvious account of our knowledge of other minds is that it rests on an argument from analogy (which belongs to the causal inference type of approach listed above). The trouble is that the obviously respectable cases of argument from analogy are where the conclusion concerns the existence of something of the same general kind as things we have encountered more directly (say that there may be life on some distant planet sufficiently analogous to our own). Ayer is still troubled, in effect, by a verificationist scruple about anything in principle unobservable.

At one stage Ayer suggested an intriguing solution. (See Ayer 1954: essay 8, also 1956: 247–9.) When I say of another that he is having a certain experience, what I mean can be analyzed into a statement of the form "A person of such and such a description (e.g. presently standing in a certain position, female, capable in philosophical argument, etc., etc.) is having such and such an experience." Now is it, Ayer asks us, a necessary truth that it is not I myself who answer to that description? If not, and it is conceivable, though profoundly contrary to fact, that I might have done so, then it is also conceivable that I could have verified the proposition directly. (So another's experience becomes more like the other side of the moon – as it seemed to be then – than, say, God.)

There is something that will seem to most people rather suspicious about this solution of the difficulty, but I cannot pursue the matter further. What it does show is that Ayer was never quite sure whether he continued to want statements which are meaningful for me to be ones which in principle I myself could verify, or whether it is only required that some human being, or like creature, could verify them.

On all these topics Ayer developed his position further. *The Central Questions of Philosophy*, in particular, includes some quite novel suggestions on our construction of the physical world, which, unfortunately, we cannot consider here.

Nor is there space here to consider the many other philosophical problems Ayer dealt with in his work. His discussion of probability, for example, does well to insist on some easily overlooked facts about it, e.g. that if statements of probability can only assert how probable something is relative to certain specifiable evidence then there is no way in which we can assert that it is more probable that a proposition based on more comprehensive evidence will be true than will one based on less comprehensive evidence.

(See 1963: 188–98 and 1972: 54–8.) His later more positive view of metaphysics, as conceptual clarification which may improve our way of understanding things, should also be mentioned. (See 1967: essay 5.)

Altogether his philosophy, right or wrong, has admirable qualities to which this account, concentrating as it does, for historical reasons, on *Language, Truth and Logic*, may not have done justice. For one thing, he wrote in philosophical prose of unrivalled excellence. It is thoroughly straightforward and extremely lucid. A passage once understood stays easily in the mind as the basis for what follows as one reads on; there was no need for those irritating numbered propositions at which one must be for ever looking back in much philosophy written today.

Ayer is for now the last great figure in the great tradition of British empiricist philosophy in the line of Hume. Its faults, so far as it is faulty (a matter not considered here), are those of that whole tradition which he may, indeed, have carried forward as well as it ever will be again.

Notes

1 I shall mostly follow the usage recommended in the introduction to the second edition of *Language, Truth and Logic*, according to which any set of synonymous declarative (he says "indicative") sentences is spoken of as expressing a statement, and this statement is said to be a proposition if and only if it is literally meaningful (1946: 8).

2 It is not altogether clear what an experiential statement or proposition is. (Ayer gives no examples here.) Does it report a present experience or observation or does it predict one expected to take place shortly? Ayer's formulation suggests the first, in which case it is rather pointless. Yet that seems to be what Ayer has in mind. In any case, experiential propositions seem very like the ostensive propositions which Ayer had rejected in the original first edition text (1946: 91–3).

3 So much, indeed, is said by Ayer himself, though somewhat as an aside. "For it will be shown that all propositions which have factual content are empirical hypotheses; and that the function of an empirical hypothesis is to provide a rule for the anticipation of experience" (1946: 41, see also p. 151 and *passim*).

It is worth noting that whereas the somewhat similarly minded American pragmatists emphasized the importance of genuine factual knowledge as facilitating our *control* of things, Ayer almost exclusively speaks of prediction and I remember him once arguing (in a seminar) against C. I. Lewis's claim that only an agent, with some control over events, could understand factual statements. (For a rare use of "control" see 1946: 50.)

4 It seems that we must either recognize a non-conventional necessity here, or agree with the idea many find in Wittgenstein that somehow each deduction from a rule is itself ultimately a free decision, or at least one necessary only in the sense of being socially enforced.

5 Compare Stevenson 1945: 68–70. The question was sometimes raised whether the verification principle was synthetic or analytic. Ayer's answer was that it was analytic with reference to the meaning which gives what we call factual statements their point. See introduction, 1946: 15–16.

6 It does not follow that physical things cannot exist unperceived, since their existence consists not in their being perceived but in the fact that if certain conditions were fulfilled they would be. Thus Ayer thinks to distinguish himself from idealism. See 1946: 145.

Bibliography

Works by Ayer

1946: *Language, Truth and Logic*, 2nd edn., London: Victor Gollancz. (First published 1936.)
1954: *Philosophical Essays*, London: Macmillan.
1956: *The Problem of Knowledge*, London: Macmillan.
1963: *The Concept of a Person*, London: Macmillan.
1967: *Metaphysics and Common Sense*, London: Macmillan.
1972: *Probability and Evidence*, London: Macmillan.
1973: *The Central Questions of Philosophy*, London: Weidenfeld and Nicolson. (Reprinted Penguin Books, 1976.)

Works by other authors

Church, A. (1949) Review of *Language, Truth and Logic*, *Journal of Symbolic Logic* 14, pp. 52–7.
Hempel, C. G. (1953) "The Empiricist Criterion of Meaning," in *Logical Positivism*, ed. A. J. Ayer, Glencoe, IL: Free Press, pp. 115–16.
Schlick, M. (1949) "Meaning and Verification," in *Readings in Philosophical Analysis*, ed. H. Feigl and W. Sellars, New York: Appleton-Century-Crofts, p. 158. (First published in *Philosophical Review* 45 (1936).)
Stevenson, C. L. (1945) *Ethics and Language*, New Haven, CT: Yale University Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

17

J. L. Austin (1911–1960)

JOHN R. SEARLE

John Langshaw Austin received his university education in classics at Balliol College Oxford. After completing his degree in 1933 he became a fellow of All Souls College and in 1935 a fellow of Magdalen College. During the Second World War, from 1939 to 1945, he served as an officer in British intelligence, rising to the rank of Lt. Colonel. He is said to be largely responsible for the extraordinary accuracy of the Allied intelligence at the time of the Normandy invasion, and he received citations from the British, French, and American governments for his war work. After the war he returned to Oxford and in 1952 he became White's Professor of Moral Philosophy, a post he held until his death in 1960.

When Austin was professor, there were about sixty practicing professional philosophers in Oxford, and only three held the rank of professor (the other two were Gilbert Ryle and H. H. Price). Austin was the most influential of a very distinguished group of Oxford philosophers of that period. During the fifties most people in Oxford thought it was the best university in the world for the study and practice of philosophy, and there was no question that philosophy was the dominant subject in the university at large. It is hard for people educated in other universities, even in Britain, to imagine the status, prestige, and intellectual centrality accorded to philosophy in Oxford at that time.

The period of Austin's ascendancy matched closely my own stay in Oxford, from my entry as a freshman in 1952 until, as a lecturer at Christ Church, I left in 1959. I got to know him quite well, and these remarks are based in part on my own personal recollections. Austin's influence was not primarily due to his writing. He published only seven articles in his lifetime, and of these only one, "Other Minds," can be said to have been tremendously influential at the time, though three others, "Truth," "A Plea for Excuses," and "Ifs and Cans" received a good deal of attention. There is a sense in which most of Austin's published works during his lifetime were popularizations. These were articles and lectures to meet some particular request or demand. Four of his articles were prepared as invited contributions to symposia of the Aristotelian Society and one was his public lecture to the British Academy. A sixth was his presidential address to the Aristotelian Society. Only one, "How to Talk – Some Simple Ways" was, so to speak, unprovoked. It was published as a separate article by the Aristotelian Society.

During his lifetime Austin's influence was due primarily to two factors: first, he had an original conception of how philosophy might be practiced; and second, he had a forceful intellect and personality that he exhibited in his teaching, and above all in philosophical discussions, both with students and colleagues. His lectures on speech acts were published after his death, and this work is his greatest legacy, though it was largely unknown in his lifetime, except to people who had been his students. Now that he has been dead for several decades, we can appraise his contributions from a longer perspective. It seems to me there are four different subjects that need to be discussed. First, his theory of speech acts. Second, his conception of ordinary language, and of ordinary language philosophy, and how it might be used constructively to give us greater philosophical insight. On this topic the classic work is his article "A Plea For Excuses." Third, Austin's conception of ordinary language philosophy and how it might be used critically in the examination of traditional philosophical issues. Austin's criticism of sense-data theories of perception, in his posthumously published book *Sense and Sensibilia*, is the purest expression of his critical technique. Fourth, much of Austin's influence both on his contemporaries, when he was alive, and on the subsequent work of his students and colleagues, was due to his qualities of character and intellect. I conclude the chapter by giving a brief assessment of his principal achievements.

The theory of speech acts

I believe Austin's most important contribution to the history of philosophy is in his overall philosophy of language as manifested in his theory of speech acts.

During his lifetime Austin's most important discovery was supposed to be that of "performative utterances," and correspondingly of performative verbs and performative sentences. In the period in which Austin worked, philosophers generally supposed that the main function of language was to make truth claims. There were various ways of describing these, and it was common to say, as the logical positivists did, that all of our cognitively meaningful utterances divided into the analytic and the synthetic, and it was common in ethical philosophy to insist that there was a distinction between those utterances which were "descriptive," and those which were "evaluative." Austin thought that all of these simple distinctions were much too crude. He was the first philosopher to notice that there is an important class of utterances made with indicative sentences that do not set out to be true or false, because in these utterances the speaker is not describing a situation, but rather performing an action, and performing an action where the utterance of the sentence constitutes the performance of the action named by the main verb of the sentence. So if I say "I promise to come and see you," in appropriate circumstances, I am not describing a promise, I am making a promise. According to Austin, the utterance of the sentence serves to perform an action, not to describe anything. This led him to make a distinction that he thought would enable us to see matters more correctly: the distinction between performative and constative utterances. There are three ways in which performatives differ from constatives. First, performatives such as "I promise to come and see you" typically have a special verb for performing the action in question, and there is even a special adverb, "hereby," which we can insert in performative sentences; for example, "I hereby promise to come and see you". Constatives, for example, "It is raining," or "Snow is white," do not have or

need a special verb. Second, constatives can be true or false, but performatives are not true or false, rather they are either felicitously or infelicitously performed. Corresponding to the true/false dimension for assessing constatives is the felicitous/infelicitous dimension for assessing performatives. And third, the performative utterance is an action, a doing, whereas the constative is a statement or a description.

However, Austin's patient research eventually showed that this way of making the distinction does not work. It turns out that all of the features that are supposed to be special to the performative are true of the constative as well, and thus what was originally supposed to be the special case, performatives, seems to swallow the general case, constatives, which now turn out to be performances of actions like any other utterance, and this led Austin to a general theory of speech acts.

Going through the three criteria in order: first, just as there are performative verbs for promising, ordering, and apologizing, so also there are performative verbs for stating, claiming, and other constatives. Thus, just as one can promise by saying "I promise," so one can state that it is raining by saying "I state that it is raining," and the criterion that Austin had hoped to use to identify performative verbs, namely the possible occurrence of the adverb "hereby," as in "I hereby promise to come," also characterizes constatives, as in "I hereby state that it is raining."

Second, the so-called constatives also have a felicitous/infelicitous dimension of assessment, and many so-called performatives can be appraised as true or false. For example, if I make a statement that I am no position to make, my utterance will be infelicitous in exactly the same sense that a promise can be infelicitous if, for example, I am unable to do the thing I promised to do. Suppose I say right now, "There are exactly thirty-five people in the next room," when I have no basis whatever for making that statement, then the statement is infelicitous in the same sense in which performatives can be infelicitous. Furthermore, there clearly are apparent performatives that can be judged as true or false. If I say, "I warn you that the bull is about to charge," when it is not the case that the bull is about to charge, then I have issued a false warning, even though a warning is a performative on Austin's original definition.

Third, making a statement is just as much performing an action as making a promise. At the end of Austin's discussion the conclusion is obvious: we should think of every utterance as the performance of a speech act. The notion of a performative should be restricted to those utterances containing the performative use of a performative expression.

The theory of speech acts begins with the rejection of the performative/constative distinction. Within the theory of speech acts Austin then made a distinction between three different levels of description of an utterance: (1) the level of the locutionary act, which is defined as uttering words with a certain meaning, where "meaning" is explained as sense and reference; (2) the level of the illocutionary act, which is defined as the utterance of words with a certain force, which Austin baptized as "illocutionary force"; and (3) the perlocutionary act, which is defined as the production of certain sorts of effects on the hearer. To take Austin's example, if I say "Shoot her," then if by "shoot" I mean shoot, and by "her" I refer to her, then I will have performed a certain locutionary act of saying "Shoot her." But if I uttered that sentence with the force of an *order*, or *advice*, or *request*, then those verbs will name the illocutionary force of my utterance and hence the illocutionary act that I was performing in making the utter-

ance. And if I *persuade* the hearer to shoot her, persuading is the production of an effect on a hearer of a sort that Austin called a perlocutionary act.

The distinction between the illocutionary and the perlocutionary seems to me essential for any theory of language, and it is especially important for those theories that take language as a matter of linguistic behavior; because, of course, the linguistic behavior which involves producing effects on people in the form of perlocutionary effects, needs to be distinguished from the linguistic behavior which involves performing speech acts, regardless of the subsequent effects on the hearers. Implicit in Austin's work is the conception that the illocutionary act, not the perlocutionary act, is the fundamental target of analysis in the philosophy of language. That has been the assumption on which I and a large number of other researchers have proceeded in attempting to carry on Austin's pioneering efforts (see SEARLE).

The distinction between the locutionary and the illocutionary, however, does not seem to me to work. The reason is that the meaning of the sentence, which is supposed to determine the locutionary act, is already sufficient to fix a certain range of illocutionary forces. You cannot distinguish between meaning and force, because force is already part of the meaning of the sentence. There is no way that I can utter the sentence "It is raining," or for that matter, "Shoot her," without performing some illocutionary act insofar as it is a locutionary act. There is no distinction between the locutionary and the illocutionary, because the locutionary is *eo ipso* illocutionary.

Austin also gave us a taxonomy of types of illocutionary acts. His taxonomy includes the following; first, verdictives. These are findings of fact or value on some matter. An example of a verdictive is giving a verdict. Second, exercitives. These are the exercising of powers, rights, and influence. Examples would be appointing and voting. Third, commissives. These are always cases of committing the speaker to a course of action. The favorite example, of course, is promising. Fourth, behabitives. These have to do with social behavior. Examples are apologizing and congratulating. Fifth, expositives. These make plain how our utterances fit into the discourse. Examples are replying, arguing, and conceding.

As with the locutionary/illocutionary distinction, it seemed to me this taxonomy needs revision and extension, because there is no clear criterion for distinguishing between the various categories. I and several other philosophers have attempted to criticize and improve on Austin's taxonomy; however it is important to emphasize that the criticisms and revisions of his views are made within a framework he invented and using tools he gave us. I see the many criticisms of Austin's specific doctrines by subsequent speech act theorists not as refutations but as further contributions to a discussion that he began, but did not live to complete.

It is important to emphasize that when we read Austin's most famous book, the work posthumously published, *How To Do Things With Words*, we are reading his lecture notes. Austin would never have published this material in this form. I know this for a fact because I wanted him to publish it so that I could publish my criticisms of it, even when I was a student. I once asked him "How soon can we hope that your William James lectures will be published?" thus giving him an opening I should never have done. He responded immediately, "You can *hope* it will be published any time you like." Further discussion revealed that he did not think the work ready for publication: "It is too half-baked," he said.

Ordinary language philosophy: the constructive function

Austin was most famous during his lifetime, not for his theory of speech acts, about which only the theory of performatives was generally known, but rather for his particular conception of philosophy, and his style of doing philosophy. He was always anxious to insist that he did not think that his was the only way of doing philosophy, but merely that it was one possible way of doing one part of philosophy. He thought that the first step to be taken in philosophy was to make a very careful analysis of the ordinary use of expressions. The ordinary expressions of a natural language like English, he thought, embodied all the distinctions about the world that people had found it necessary and useful to make in the course of millennia. He did not think that ordinary language was the last word, but he did think it was the first word. In a debate with Bertrand Russell, when Russell asked him if he thought the examination of ordinary language was the be-all and the end-all of philosophy, Austin is reported to have answered, "It may not be the be-all and the end-all, but it certainly is the begin-all." The analysis of the ordinary use of expressions served two philosophical purposes for Austin. One was a corrective purpose of showing that many of the claims that philosophers had made rested simply on mistakes about the ordinary use of expressions. His most famous discussion in this regard is probably his criticism of the arguments for the sense-data theory of perception, in his lectures *Sense and Sensibilia*. The second purpose of the analysis of language was more constructive: he thought we could learn a great deal about the world by analyzing the expressions we use to describe the world.

Austin thought that his method of doing philosophy allowed for two features which philosophy is thought not to possess. First, philosophy on his conception is a cooperative enterprise. It is not something you do alone in your study, but rather you get a group of people and try to discuss examples to see how words are used in describing those examples. And second, philosophy so construed allows for progress. It is typical that people who carry on philosophical discussion of this type, at the end of the day, feel they have made definite progress in analyzing the application of words to concrete examples.

It is characteristic of Austin's approach that he can often show that what seem like two synonyms or near synonyms are really quite different. In a famous case he took the two expressions "by accident" and "by mistake" and showed that they really had quite different meanings, even though at first sight most English speakers would probably say they mean pretty much the same thing. Here was his demonstration. Suppose I go out into the field to shoot my donkey. Suppose I see your donkey, which looks very much like my donkey, and I shoot your donkey. Did I shoot your donkey by accident, or by mistake? Suppose I go out into the field and shoot at my donkey, but just as I am pulling the trigger, the two beasts move, and my bullet strikes your donkey. Did I shoot your donkey by accident, or by mistake? I think the examples are absolutely clear in both cases.

The constructive side of Austin's method of doing philosophy is most powerfully exemplified by his article "A Plea For Excuses" (together with its posthumously published companion piece, "Three Ways of Spilling Ink"). "A Plea for Excuses" is, in fact, a summary of an entire series of seminars that Austin gave during the 1950s. I

attended the seminars, and there was easily enough material presented to fill an entire book, but it is perhaps typical of Austin that the material that a more average philosopher would use for a complete book he condensed into a single article. Austin's method is illustrated by the following: most philosophers, myself for example, if examining the problem of action, would begin by asking what fact about an event makes it into a human action. Austin thinks that this approach, as he frequently said in criticizing my views, is "much too fast." His own approach is, so to speak, to sneak up on the problem indirectly by asking what sorts of excuses, justifications, extenuations, and explanations we offer for our actions. "Excuses," he insists, is just a title, not a description of the whole subject matter. The results of the analysis are a series of theses that he advances about the character of our conceptual apparatus for discussing actions. Many of these are quite surprising. So, for example, I think most philosophers intuitively would suppose that any action is done either voluntarily or involuntarily. But Austin points out that the whole question of negations and opposites is much more complex than that. The opposites of the word "voluntarily," he says, might be "under constraint" or "duress" or "obligation." The opposite of "involuntarily" might be "deliberately" or "on purpose" or the like. Austin urges us not to take anything for granted about negations and opposites. Again, I think many philosophers suppose that there is not much difference between doing something intentionally, deliberately, and on purpose. Austin makes it abundantly clear that these are not at all the same. He also urges us to pay close attention to legal cases and psychological studies; he examines one case, *Regina* v. *Finney*, in some detail, showing that the lawyers and the judge make serious mistakes, treating several terms of excuse as equivalent when they are not, and being unclear about what action exactly of the defendant is being qualified by what expression. It is impossible to summarize this article because the article is itself a summary of a quite extended project of research, and the interest of the results is in the specific details. But the article reveals both the strengths and some of the limitations of Austin's method.

Ordinary language philosophy: the critical function

I believe the purest case where one can observe Austin, so to speak "in action," is in his book *Sense and Sensibilia*. The actual text that we have before us now was prepared from notes of numerous students by Geoffrey Warnock, but Warnock does an excellent job of conveying the flavor of the actual lectures, as I can say from having attended them. If Austin had lived, I doubt that he would ever have published these lectures as they stand. Their results are almost uniformly negative, and the tone is often more harsh than Austin would normally have allowed in publication. Nonetheless, they are a beautiful exemplification of his method of philosophical analysis. He simply does a careful word-by-word examination of a series of traditional philosophical arguments designed to show that we never perceive "material objects," but only perceive "sense-data." Austin takes Ayer's book *The Foundations of Empirical Knowledge* as his "stalking horse," and he also discusses arguments from Price and Warnock (see AYER). Austin goes patiently through the arguments that are traditionally called "the argument from illusion," which attempt to prove that all we ever perceive are sense-data, and he shows that without exception the arguments, as presented by Ayer, are hopelessly muddled

and confused. Ayer assumes that such words as "look," "appear," and "seems" can be used indifferently as if they meant the same thing, but Austin's patient analysis shows that they are really quite different. In the standard arguments for sense-data Austin finds only carelessness, muddle, and confusion.

The stages of the arguments that he finds are

1 The philosophers assume that there are two exclusive classes of sense experience, those of material objects, and those of sense-data.
2 They argue that there must be no discriminable differences in the character of the perceptions since we can confuse one thing for the other.
3 They conclude that, since one would expect a considerable difference from two such different sorts of entities, there must be only one class that we are actually perceiving, and that one must be sense-data.

By patiently working through the texts Austin challenges each of these claims.

1 There are all sorts of things we perceive that do not fit either category comfortably, such things as shadows, clouds, gases, flames, rainbows, images, etc. Austin thinks the dichotomy is a typical philosophers' oversimplification. It would be just as confused to say that all we perceive are material objects as it is to say that all we perceive are sense-data.
2 In real life there are all sorts of differences in the character of our experiences. Dreams, for example, are different from waking experiences in all sorts of ways; and even the stock-in-trade waking-life examples of the epistemologist are misdescribed. For example, the stick in water which "looks bent" does not look like a bent stick out of water, and even in water it need not look *to be* bent.
3 It does not follow from the arguments as presented that all we ever see are sense-data, and indeed it is quite arbitrary for the philosophers to select "sense-data" or "material objects" as the objects of perception.

It is important to emphasize that in criticizing the sense-data theory, Austin is not defending the idea that all we see are material objects. He thinks that idea is just as crude as its opposite.

In the course of his discussion he introduces the idea of what he called a "trouser word," and what some philosophers subsequently came to call "excluders." Some words get their meaning in a context from the words that they are opposed to in that context. Thus *real* cream is opposed to *artificial* cream, but a *real* duck is opposed to a *toy* duck or a *decoy* duck, and *real* teeth are opposed to *false* teeth. The word "real" is an excluder that gets its meaning in context from what it is opposed to. There is no common property of reality which the word "real" invariably and literally serves to ascribe. A decoy duck for example, though not a real duck, may nonetheless be a real decoy, as opposed for example to a paper model of a decoy duck. When the epistemologist talks about reality and perceptions of reality, he fails to appreciate the nature of the concept.

In a reply to Austin, Ayer claimed that his main points could survive, even if he accepted all of Austin's specific objections. His main point is that we could have all of the experiences we do have and still be mistaken in our claims about objects and states of affairs in the world. To this I think Austin would have replied, first that this does not

show that all we ever perceive are the experiences, the "sense-data." And, second, it does not show that the relation between the experiences and the objects they are taken to be experiences of, is one of evidence. It does not show that the experiences are evidence for the presence of the objects.

Whatever the merits of the debate, I think Austin's critique proved immensely influential historically. One used to hear a lot about the sense-data theory of perception; one does not hear much about it anymore.

Other works

Austin wrote a number other important works which limitations of space prevent me from exploring in any detail, but I must mention them in passing. "Other Minds" presents a criticism of traditional epistemology which is very much in the spirit of *Sense and Sensibilia*. "Truth" and "Unfair to Facts" present a version of the correspondence theory of truth and a response to Strawson's criticisms of it. I think Austin is right that the fundamental notion of truth is correspondence, but the particular version that he presents does not survive Strawson's objections. "Ifs and Cans" is a gem of philosophical analysis. As far as the history of philosophy is concerned, its main point is to respond to those versions of compatibilism about the free will problem which maintain that to say that I could have done otherwise just means that I would have done otherwise if I had so chosen, and to say that I can do something just means that I will do it if certain other conditions are met. But Austin makes a large number of other points about related conceptual issues.

Character and intellect

My own impressions of Austin are somewhat different from those of many people who thought of themselves as his close associates and followers. One trait that we would all agree on was his immense carefulness and precision. Not only when doing philosophy, but even in the most casual conversation, Austin spoke and thought with great precision, and he did not tolerate looseness in his students or colleagues. The worst condemnation that he could make of something he was reading would be to shake his head sadly and say in his thin, precise way, "It's just loose." Indeed, on several occasions he said to me in tones more of sadness than anger, "There is a lot loose thinking in this town." For the most part, Austin's colleagues regarded him with a kind of awe, and it seemed to me that to some extent they were even terrified in his presence. Certainly, his presence in seminars and meetings had a profound effect on the behavior of others participating. I noticed this when I went back to Oxford some years after his death and found that many of the professional philosophers were behaving like schoolboys during recess. They were much less cautious than they would have been in Austin's presence.

At first I could not understand the source of Austin's influence, because it seemed to me that I could I beat him in argument; and like a lot of undergraduates I thought the test of a philosopher was how good he was in the give and take of philosophical repartee. Austin's technique in discussion was always to take everything dead literally, and then to insist on certain linguistic distinctions that he thought were being overlooked. So, for example, when Wittgenstein's *Philosophical Investigations* was published,

I and some other undergraduate philosophers insisted that we discuss Wittgenstein's private language argument in Austin's informal instruction for undergraduates. Austin's technique was to refuse to grant Wittgenstein any leeway at all. At one point when we were discussing Wittgenstein's famous example of the beetle in the box, Austin said sarcastically, "All right, for our next session everyone bring a box with a beetle in it." At one point in Wittgenstein's discussion, he says there might be nothing at all in the box. Austin thought Wittgenstein was simply contradicting himself. "First he says, there definitely is a beetle in the box, and then he says there might be nothing in the box, a plain contradiction."

Austin's habit of insisting on the highest level of precision, both in his professional activities and even in ordinary conversations, seems to me one of the main reasons for the terror that he inspired in his colleagues. At the time, much of the source of Austin's influence derived from his schoolmasterly style. Most Oxford philosophers of the time had been students at British boarding schools, and Austin was, so to speak, the ultimate schoolmaster. If he were reading a paper that I had written, a typical question he put to me would be, "Why exactly did you use the subjunctive?" Or, on another occasion, "In the verb 'suppose,' what does the 'sup' mean?" Austin was famous at the time for his attention to the minute details of ordinary language, but it seemed to me then, and it seems to me now, that his real contribution to philosophy was not so much in the details. Austin did indeed have a genius for spotting linguistic differences and distinctions where most people would have thought there were none, though in the details his views were sometimes mistaken. His most important contribution to philosophy, I believe, is in his overall vision of language.

Though he was regarded as terrifying by many of his colleagues, I can say that to undergraduates he was immensely kind, patient, helpful and, in his reserved way, even friendly. His contempt was reserved for people he thought of as pretentious, self-important, pompous, and above all obscurantist. When Austin, along with several other Oxford philosophers and students, went off to Royaumont in France for an English–French philosophy colloquium he considered the pretentiousness of Merleau-Ponty, then the most influential French philosopher, quite ridiculous. "That Merleau-Ponty, he is just a little tin god. He will never get anywhere." Austin would have hated the "deconstructionists" and "postmodernists" who currently pretend to admire his work.

He did not think his brand of philosophy was the only correct way to do the subject, but he did try to extend its influence with an almost missionary zeal. Again, when we were in Royaumont, and he saw me in discussion with an elderly distinguished French philosopher, he took me aside and said, "Don't waste your time on the aged. Talk to the young!" This remark annoyed me at the time because I did not think of myself there for any other reason than to practice philosophy. Austin, I believe, would have thought it a waste of time for us to go to France if we did not try to spread the truth.

I never heard anyone other than his wife address him by his first name, and when one of his colleagues had the temerity to address him as "John," Austin is reported to have said evenly, "'Austin' is also a Christian name." He did have a habit of holding one's attention in discussion. So, for example, when making an involved point he would take his pipe in one hand and light a match with the other. Never taking his eyes off his interlocutor, he would allow the match to burn ever closer to his bare fingers until

at the last millisecond he would flick his wrist to extinguish the match, whereupon he would start the whole process over again while continuing his relentless discourse, eyes always on his listeners.

Austin's reluctance to publish was part of the culture of Oxford at the time, but also it was partly characteristic of his own attitudes. Oxford had a long tradition of not publishing during one's lifetime, indeed it was regarded as slightly vulgar to publish. People who did publish a lot, like A. J. Ayer, were regarded as remiss for having published too much too soon. As far as having a career and making a reputation were concerned, the attitude in Oxford was that the only opinions that really matter are the opinions of people in Oxford, and perhaps a few in Cambridge and London, and they will know about one's work anyway. One does not need to publish. What one does not want is a lot of graduate students somewhere, picking over one's half-baked published texts and – horror of horrors – finding mistakes. So I think Austin's reluctance to publish was partly due to his extreme carefulness, but it was partly due to his sheer vanity; he did not want any intellectual inferior pointing out errors.

At a time when anti-Americanism was very common in Britain, especially among the intellectual classes, Austin simply adored the United States, and especially its university system. He would not tolerate criticisms of the United States, and the only subject on which I have ever heard him show uncritical enthusiasm was America. Indeed, he once said to me, "The future lies with America," and on another occasion, "There are unplowed fields in that country." Once Herbert Hart was criticizing American cooking, and Austin said evenly, in his discussion-ending way, "It is not so bad."

I often read how much Austin was influenced by Wittgenstein. Nothing could be further from the truth. Austin had no sympathy whatever for Wittgenstein, and I think he was incapable of learning from someone whose style was so "loose." He typically referred to Wittgenstein in the style of English schoolboy slang of the time as, "Witters," pronounced "Vitters." He thought there were no original ideas in Wittgenstein. Indeed he once said to me about Wittgenstein's philosophy, "It's all in Moore," one of the least accurate things I have ever heard Austin say. If Austin had an inspirational model, it was Moore (see MOORE).

It will seem a paradoxical feature of Austin's career that he aroused such passionate controversy both pro and con. On the surface, at least, his presentations are invariably modest, cautious, and self-effacing. Though I did not regard myself as one of his followers, I found it easy to see why they regarded him with such enthusiasm. He offered them a new conception of philosophy, and with it, a new research program. But it is more puzzling to try to understand why he was hated so much. I think in order to understand the hostility that he aroused, we have to compare his career with that of Socrates. He was hated for much the same reason that Socrates was hated: he seemed to destroy everything without leaving anything substantive in its place. Like Socrates he challenged orthodoxy without presenting an alternative, and equally comforting, orthodoxy. All Austin offered, again like Socrates, was a new method for doing philosophy.

Austin's substantive achievement, especially in the theory of speech acts, has survived his death now for nearly a half century, and will, I believe, continue to be a focus of research. But his official doctrine as to how philosophy might be pursued has waned

considerably. It has very few followers and practitioners. Why? Well, part of the answer is that it is just too difficult. The sort of very careful analysis of minute linguistic distinctions that Austin urged us to undertake is simply too much work for most philosophers. Austin thought that we ought to be more patient and hardworking. If entomologists can classify a million different kinds of insects, surely philosophers ought to have the patience to classify the few dozens or few hundred or even a few thousand different sorts of uses of different sorts of words. But the problem is that the motivation that tends to make one a philosopher seems to be quite different from the motivation that makes one an entomologist. Philosophers want very general answers to very large questions, whereas Austin thought they had first better get clear about the distinctions among a number of adverbs, working themselves up to undertake an analysis of a few verbs.

There are certain limitations on Austin's methods, which, paradoxically but to his credit, we can use Austin's theory of speech acts to expose.

(1) Sometimes Austin confuses the truth conditions of a term, that is, the conditions under which it is a fact that some object satisfies that term, with the conditions for appropriately *asserting* that the term applies. Thus, to take an Austin-style example, Austin points out that we wouldn't normally assert that a man walked across the room intentionally, if he just walked across the room in an ordinary, unexceptional way. "No modification without aberration," Austin tells us. Nonetheless it may be *true* that the man walked across the room intentionally, it is just not appropriate to assert it unless there is something unusual. It may just be too obvious that the act was done intentionally.

(2) Related to the first mistake is the mistake of confusing the meaning of a term with the illocutionary force that characteristically accompanies the assertion that the term applies to an object. Thus, to take another example from Austin, he points out that when we say that we know something, we are often giving our guarantee for what we claim to know, that the claim to know has certain features in common with a performative, such as "I promise." Austin is careful not to say that "know" is a performative, but he does think that the assertion that one knows has a performative-like guaranteeing force. But once again this does not tell us the meaning of the word "know," because it cannot account for the occurrence of this word in other cases such as conditionals or negations. So even if an utterance of the form "I know that *p*" means something like "I guarantee that *p*," still, an utterance of a conditional of the form "if I know that *p* then *q*," does not mean anything at all like "if I guarantee that *p*, then *q*."

(3) Even after you have done a careful linguistic analysis and shown that the standard philosophical positions rest on a misuse of words, still, you can often state the position again without using those words. The problem remains even after the misuse of words have been corrected. Thus, to take the problem of free will, Austin points out that when we say that an act was done freely, "freely" functions as an excluder, excluding all of the various ways in which an act may not have been done freely, such as, for example, done under duress or under compulsion. But even if Austin is right about this, and he probably is, you still have a free will problem left over. Here it is: Are all human acts such that the performance of the act has antecedent causal conditions which are causally sufficient to determine the act? If I walk across the room, and I walk across

the room in a way which is not under duress or compulsion, all the same there is still the question about free will. Were the antecedent conditions prior to the onset of my action of walking across the room sufficient to determine that I was going to walk across the room? That question remains even after we have become clear about all of the various uses of freely, voluntarily, etc.

(4) Austin says that ordinary language embodies all of the distinctions that humans have chosen to make over millennia. But there is a sense in which that is not quite right. We can indeed *state* in ordinary language all of the distinctions that humans have chosen to make, and indeed we can state a lot that they have not chosen to make. But it is not the case that every real distinction that humans have made is marked by a lexical distinction, by two different words, in ordinary usage. Thus to take one of Austin's examples, the word "pretend" does not mark a distinction between those cases of pretense which are genuinely intended to deceive, and those cases of pretense which are put on or are mock-performances, but not designed to deceive. So if I pretend to be the President of the United States in order to be admitted to the White House, I have pretended in the deceptive way. But if I pretend to be the President of the United States as part of a game of charades, there is no intention to deceive. This is an obvious and important distinction, as Anscombe pointed out in the symposium with Austin on pretending, but we do not have two verbs whose meanings are "pretend deceptively" and "pretend non-deceptively."

Conclusion

J. L. Austin was one of the most important philosophers of the twentieth century. In examining his contribution we need to distinguish between the philosophy of language and linguistic philosophy. The philosophy of language is the attempt to give an account of certain very general features of the structure, use, and functioning of language. Linguistic philosophy is the attempt to solve philosophical problems by using linguistic methods. Austin made important contributions to both the philosophy of language and linguistic philosophy. During his lifetime he was famous as a linguistic philosopher, but not for his philosophy of language. Since his death it has emerged that his most important contribution to philosophy has been his philosophy of language as expounded in his theory of speech acts. At the conclusion of "Ifs and Cans," Austin expresses the hope that the next century may see the birth of a comprehensive *science of language*. I believe that he thought his theory of speech acts was a contribution toward that future science.

Bibliography

Works by Austin

1962: *How To Do Things With Words*, Cambridge, MA: Harvard University Press.

1964: *Sense and Sensibilia*, Oxford: Oxford University Press.

1979: *Philosophical Papers*, 3rd edn., Oxford: Oxford University Press. (2nd edn., 1970.) Articles in this collection are given below, with first publication details where applicable:

"Agathos and Eudaimonia in the *Ethics* of Aristotle," from *Aristotle: A Collection of Critical Essays*, ed. J. M. E. Moravcsik, New York: Doubleday, 1967, and London: Macmillan, 1968.

"A Plea for Excuses," *Proceedings of the Aristotelian Society* 57 (1956–7).

"Are There *A Priori* Concepts?," *Proceedings of the Aristotelian Society*, Annual Conference, 1939.
"How to Talk," *Proceedings of the Aristotelian Society* 53 (1952–3).
"Ifs and Cans," *Proceedings of the British Academy* 42 (1956).
"Other Minds," *Proceedings of the Aristotelian Society*, suppl. vol. 20 (1946).
"Performative Utterances."
"Pretending," *Proceedings of the Aristotelian Society*, suppl. vol. 32 (1957–8).
"The Line and the Cave in Plato's Republic." (Not in 1st and 2nd edns.)
"The Meaning of a Word."
"Three Ways of Spilling Ink," *Philosophical Review* 75/4 (1966).
"Truth," *Proceedings of the Aristotelian Society*, suppl. vol. 24 (1950).
"Unfair to Facts."

Works by other authors

Anscombe, E. (1957–8) "Pretending," *Proceedings of the Aristotelian Society*, suppl. vol. 32.
Ayer, A. J. (1940) *Foundations of Empirical Knowledge*, London: Macmillan.
——(1973) "Has Austin Refuted Sense-data?," in *Essays on John L. Austin*, ed. I. Berlin et al., Oxford: Clarendon Press.
Fann, K. (ed.) (1969) *Symposium on J. L. Austin*, London: Routledge and Kegan Paul.
Holdcraft, D. (1978) *Words and Deeds*, Oxford: Oxford University Press.
Searle, J. (1969) *Speech Acts: An Essay in the Philosophy of Language*, Cambridge: Cambridge University Press.
——(1979) *Expression and Meaning: Essays in the Theory of Speech Acts*, Cambridge: Cambridge University Press.
——(1989) "How Performatives Work," *Linguistics and Philosophy* 12, pp. 535–58.
Travis, C. (1975) *Saying and Understanding*, Oxford: Blackwell Publishers.
Warnock, G. (1991) *J. L. Austin*, Arguments of the Philosophers series, New York: Routledge.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

18

Norman Malcolm (1911–1990)

CARL GINET

Introduction

Norman Malcolm was born on June 11, 1911, in Selden, Kansas, and died in London on August 4, 1990. His undergraduate years were at the University of Nebraska, where O. K. Bouwsma was one of his teachers. His Ph.D., granted in 1940, was from Harvard, but the most important philosophical influences on him during his graduate years were G. E. Moore and Ludwig Wittgenstein, with whom he studied during a fellowship at Cambridge University in 1938–9. He was an instructor at Princeton before joining the US Navy in 1941. After the war he spent another year in Cambridge, 1946–7, studying with Moore and Wittgenstein. In 1947 he joined the Sage School of Philosophy at Cornell, where he remained until his retirement in 1978. During the last twelve years of his life he lived in London and was appointed a Visiting Professor and Fellow at King's College London, where he gave a weekly seminar mainly devoted to the philosophy of Wittgenstein.

Malcolm credited Moore with being the first to employ the technique of refuting paradoxical philosophical statements by pointing out that they go against ordinary language – that they imply that ordinary uses of language are incorrect uses – which is Malcolm's own favorite technique. From Wittgenstein he took the idea that a philosophical problem is essentially a confusion in our thinking that is to be remedied by reminders of the actual use of language, and by reconstructing and criticizing the analogies and reasoning that bewitch the victim of the puzzle.

Malcolm was a major expounder and endorser of Wittgenstein's later philosophy. He devoted several articles explicitly to explaining Wittgenstein's thought. The earliest and probably most influential of these was his discussion of Wittgenstein's *Philosophical Investigations* published in *The Philosophical Review* in 1954, which prompted a good deal of interest in Wittgenstein's argument against the possibility of a private language. Wittgenstein visited Malcolm in Ithaca in 1949 and their discussions there of knowledge and certainty stimulated the thinking that led Wittgenstein to his last major work, *On Certainty*. Late in his life, in 1986, Malcolm published a book, *Nothing is Hidden*, the aim of which is to expound Wittgenstein's later criticism of his *Tractatus Logico-Philosophicus*. And at his death he left a monograph, *From a Religious Point of View?* (published posthumously, edited and with a response by Peter Winch) in which he

summarizes much of Wittgenstein's work, early and late, in an attempt to see in what sense Wittgenstein's remark that he approached every problem from a religious point of view might be true. But nearly all of Malcolm's work, from the early 1950s on, is shot through with approving reference to remarks of Wittgenstein's. It would not be an exaggeration to say that he aimed nearly all of his work at getting across insights he owed to Wittgenstein (see WITTGENSTEIN).

Malcolm's writing is remarkable for its clarity and vigor and its freedom from technical jargon. It is crammed with down-to-earth examples. These sometimes help to give a concrete grasp of an abstract idea, but typically they serve to remind his readers of how ordinary language is actually used. They always give his writing considerable charm.

What follows are brief expositions of some of the views Malcolm argued for (I find no evidence in his writings of any major change in his views), placed under four headings: knowledge, mind, memory, and philosophy of religion. This sample is far from comprehensive, but I hope it gives a good idea of the breadth and character of Malcolm's work.

Knowledge

In two important early papers, "Certainty and Empirical Statements" (1942) and "The Verification Argument" (1950), Malcolm rebutted the claim made by some philosophers (e.g. C. I. Lewis, Carnap, Russell, Ayer) that it is impossible for an empirical statement (a contingent statement about material objects) to be known with certainty. In Lewis and Carnap he finds an argument for this paradoxical claim, which he calls the Verification Argument. He arrives at the following formulation of the argument ("The Verification Argument", in 1963a: 26):

I. Any empirical statement *S* has consequences (not in the sense of entailment but in the sense in which it is a consequence of "Yesterday the phrase 'the stream of thought' was on page 224 of vol. I of my copy of James's *The Principles of Psychology*" that if I were to look on that page now I would see that phrase).
II. The consequences of *S* are infinite in number.
IIIa. It is not certain that the consequences of *S* will occur.
IVb. If any empirical statement can be conclusively established as true or false, then if a sufficient number of the consequences of *S* should fail to occur then it would be absolutely conclusive that *S* is false.
Va. If at any time it should be absolutely conclusive that *S* is false then at no previous time did anyone make absolutely certain that S is true.

It does follow from these premises that, for any empirical statement S, no one ever made absolutely certain that *S* is true. Malcolm denies premise IIIa. He argues that it was accepted by proponents of the Verification Argument only because they thought it follows from III: It is possible that the consequences of *S* will fail to occur. But they thought this only because they failed to distinguish among different interpretations of III. For those senses of "possible" in which III is true ("The consequences of *S* will fail to occur" is not self-contradictory; no consequence of *S* is entailed by the grounds for holding it true), III does not entail IIIa; and for those senses of "possible" in which

III entails IIIa (there is some reason to believe that the consequences of *S* will not occur; there is no reason to think that the consequences of *S* will occur; the grounds for holding that the consequences of *S* will occur are not absolutely conclusive), III is not true.

In his earlier work on knowledge, Malcolm seemed to share the common assumption that statements of the form "*S* knows that *p*" state a fact about *S*, a fact about whose necessary components philosophers might hope to say something informative, e.g., that *S* must believe *p*, that *p* must be true. But in many of his later writings, Malcolm seems to treat such sentences, particularly in the first-person present-tense form, as like performatives whose use is, not primarily to report a fact about the subject, one that is there independently of any utterance, but rather to achieve some aim of the speaker, so that the correctness and intelligibility of its use depend heavily on the context of the use. (There were, however, already intimations of this idea in early papers. In "Defending Common Sense" (1949) he said that Moore misused "know" in making such assertions as "I know that I am a human being" or, when holding up his hand, "I know that this is a hand," in the contexts in which he made them because there was not any doubt or disagreement about the matter that would give a point to such assertions. In "Philosophy for Philosophers" (1951), for such reasons as that a sentence like "I know I feel hot" is almost never seriously used, that the normal usage of "I know" is informative and connected up with investigating, finding out, making sure, producing evidence, with asking and answering "How do you know?", he said that "In the sense of 'knowledge' in which knowledge is contrasted with belief, we do not (and cannot) have knowledge of our own sensations" (p. 336).)

His 1976 paper "Moore and Wittgenstein on the Sense of 'I Know'," says that "I know" does a variety of jobs in ordinary language use; for example, "it is used to claim the possession of evidence, or expertise, or ability; it is used to comfort, reassure, express agreement; it is used to say that one has thoroughly checked something, or that one can be relied on, or that one doesn't need to be reminded" (1977b: 192). For such reasons he often says that, except in very special contexts, it makes no sense to say "S knows that he is in pain" and what it means when it does make sense is very different from what is meant by "S knows that there is a gash in his hand." He refuses to accept that there is in this case any distinction (such as has been suggested by Grice, Searle, and others) between truth-conditions and requirements for the aptness of asserting (see GRICE and SEARLE). It is unclear what his view would be about whether there is such a distinction for such sentences as "Moore is a human being" or "What Moore is holding up is a hand."

Mind

There are two connected principles about psychological concepts that are fundamental for Malcolm. One is: Only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious . . . thinks (see Wittgenstein, *Philosophical Investigations* §282, §360). We ascribe mental properties to others on the basis of observable behavioral criteria that are non-contingently connected to the concepts of those properties; it is part of having the concepts of the properties to know behavioral

criteria that justify ascribing the properties in the right circumstances. The second is: It is not on the basis of any criteria that we ascribe mental properties (current, conscious ones) to ourselves; our self-ascriptions are analogous to, and in some cases simply replace, natural manifestations of mental states, such as the expression of pain in crying or moaning; they serve others as criteria for ascribing the properties to us. "First person utterances, and their second and third person counterparts," he says, "are linked in meaning by virtue of being tied, in different ways, to the same behavioral criteria" (1971: 91). Cartesian philosophy of mind, which he sometimes calls "introspectionism," runs afoul of the first principle. Behaviorism as a philosophy of mind runs afoul of the second.

Against the thesis that mental states or processes are identical with brain states or processes Malcolm marshals several different arguments. In "Scientific Materialism and the Identity Theory" (1964) he argues against J. J. C. Smart's claim that a sudden thought is contingently identical with a brain process as follows: we attach no meaning to determining the bodily location of a thought; so, if x is identical with y only if x and y occur at the same place and time, and the identity is contingent, then there can be no way of establishing that this same location condition is satisfied. (In the same paper he remarks that the senselessness of the supposition that a separated brain could have thoughts or sensations "seems so obvious that I find it hard to take it seriously" (p. 124).) In "Functionalism in Philosophy of Psychology" (1980) he imagines Mr. A saying to his wife, "Are you always on time?" and argues that one knows that Mr. A meant his utterance sarcastically and not admiringly only by knowing something of the previous course of their lives together, so that there is no way in which the presence of the one intention rather than the other can be accounted for by some story about neural firings or electric potentials within Mr. A at the time of his utterance. In *Consciousness and Causality* (1984) he argues that mental states without genuine duration (abilities, dispositions, intentions, beliefs) cannot be identical with brain states which do have genuine duration, and he argues that, since having an intention with a certain content entails having the concepts required to understand that content, it is impossible to identify the intention with a brain state, because possession of those concepts would, presumably be identified with other brain states and it is only contingent that they occur in the same brain.

In his "The Conceivability of Mechanism" (1968), which has been much cited in subsequent discussions of mental causation, Malcolm argues that a completely mechanistic explanation of a piece of human behavior – one entirely in terms of physical states and processes in the organism – is incompatible with any intentional or purposive explanation of it. He finds untenable both of the two ways he sees of trying to maintain their compatibility: maintaining that intentional concepts can be defined in terms of non-intentionally specified behavioral dispositions and maintaining that intentional states or events are contingently identical with neurophysiological states or events. If all human behavior had sufficient mechanistic causes then, he argues, human beings would have no intentions or desires. And, he observes, there would therefore be a pragmatic paradox in anyone's asserting that all human behavior is mechanistically explicable: since the asserter's utterance could count as an assertion only if he has certain intentions about it, his asserting this would constitute a counterexample to what he asserts.

Inspired by a remark of Wittgenstein's about dreaming, Malcolm notoriously argued, in a paper "Dreaming and Skepticism" (1956) and a book *Dreaming* (1959), that dreams cannot take place during sound sleep, in the sense of occurring at definite times and having definite durations. He infers this (and the stronger conclusion that there can be no mental activity during sound sleep) from the premise that the concept of sound sleep precludes the subject's manifesting any mental activity while sound asleep. He says that, if we found a correlation between some physiological process during sleep and reports on awaking of dreams and used that as a basis for locating dreams in objective time, that would be to adopt a different use of "dreaming" than we now have, a new meaning for the term. "As things are," he says, "the notions of duration and time of occurrence have no application in ordinary discourse to dreams. In this sense, a dream is not an 'occurrence' and therefore not an occurrence during sleep" (1956: 30).

Malcolm made a significant contribution to the study of Descartes's philosophy of mind in two papers, "Descartes' Proof that His Essence is Thinking" (1965) and "Descartes' Proof that He is Essentially a Non-material Thing" (1975). The first conjectures that Descartes argues as follows: "*x* is my essence if it is the case that (a) if I am aware of *x* then (necessarily) I am aware of myself, and (b) if I am aware of myself then (necessarily) I am aware of *x*. Thinking satisfies these conditions. Ergo, thinking is my essence" (1977b: 32). This argument, Malcolm suggests, could be Descartes's reason for thinking that he has a clear and distinct idea of himself as a thing with no corporeal characteristics. Malcolm's criticism of the argument is that, although (a) and (b) are true when "thinking" is substituted for *x*, this is not because of any necessary connection between myself and thinking. He says: "(a) is true solely because the statement 'I am not aware of myself' is self-defeating . . . (b) is true because the awareness of anything is thinking, and also because of Descartes' doctrine that one cannot think without being aware of thinking" (1977b: 36).

In the second paper, Malcolm, responding to a suggestion from Robert Jaeger, finds textual support in Descartes for the following argument: "I think I am breathing entails I exist. I think I am breathing does not entail I have a body. Therefore, I exist does not entail I have a body." Malcolm rejects the second premise. It is, he says, conceptually impossible for me to exist without ever having had a body, or for minds to exist without there ever having been bodies, because the primary use of "He thinks he is breathing" presupposes behavioral criteria of its truth (and secondary uses in speaking of ghosts or disembodied deities could not exist without primary uses). He points out that Descartes could be hoist with his own petard here, for I am breathing entails I exist but does not entail I am thinking.

In "Thoughtless Brutes" (1972 presidential address to the Eastern Division of the American Philosophical Association) Malcolm argues that the reason Descartes claims that animals do not have "real" sensations is that he insists that "when we mean by 'sensation' something other than mere physiological processes, then sensation [has] propositional content" and he thinks, rightly, that propositional representations do not occur in the "lower" animals. Malcolm comments, "When we see the enormity of [Descartes's] exaggeration of the propositional in human life, our unwillingness to ascribe propositional thinking to animals ought no longer to make us refuse to attribute to them a panoply of forms of feeling, of perception, of realization, of recognition, that

are, more often than not, nonpropositional in the human case" (1972: 53). He adds, "We need to avoid identifying thoughts with their linguistic expression. At the same time we should reject the suggestion that it is possible that language-less creatures should have thoughts . . . [F]or it is meaningful to suppose that a person might have had a thought to which he gave no expression, only because this person speaks or spoke a language in which there is an institution of testifying to previously unexpressed thoughts" (p. 55).

Memory

Malcolm's work on memory is found in his "Three Lectures on Memory" (1963b) and a book, *Memory and Mind* (1977a). In the first lecture, "Memory and the Past," he argues against Russell that the hypothesis that the world began five minutes ago complete with misleading records, delusory memories, etc., is not "logically tenable." His main argument is that a linguistic community can be said to have mastered the past tense, and therefore make past tense statements and have past tense beliefs, only if not all of their past tense statements are false. He also asserts that, if our apparent memories largely agree with each other and with the records then the apparent memories would be verified as true, and "if the apparent memories were verified it would not be intelligible to hold that, nevertheless, the past they describe may not have existed" (1963a: 199).

In the second lecture, "Three Forms of Memory," he distinguishes factual memory (remembering that *p*), personal memory (remembering something one previously perceived or experienced), and perceptual memory (personally remembering something by forming a mental image of it). He says that while a personal or perceptual memory always entails some factual memory, there can be a factual memory that does not entail any perceptual or personal memory (contrary to Russell and others). There could, he says, be a person who lacked perceptual memory altogether but had more or less normal factual and personal memories, but there could not be a creature we would recognize as a human being who altogether lacked factual or personal memory.

In the third lecture, "A Definition of Factual Memory," he suggests the following definition: "A person, *B*, remembers that *p* from a time, *t*, if and only if *B* knows that *p*, and *B* knew that *p* at *t*, and if *B* had not known at *t* that *p* he would not now know that *p*" (1963a: 236). Concerning the third, counterfactual conjunct here, he says, "Whether or not it makes sense to postulate a specific brain-state or neural process persisting between the previous and the present knowledge that *p*, such a postulation is obviously not required by an analysis of the concept of remembering," and guesses "that our strong desire for a mechanism of memory arises from an abhorrence of the notion of action at a distance-in-time" (1963a: 237–8).

In the book he maintains that it is an error to think that the causal ingredient in memory requires the assumption either of a temporally continuous chain of causation or of causal laws. He argues against the idea that there must be a representation in remembering and the idea that there must be a structural isomorphism between an occurrent memory, what is remembered, and an intervening brain state or process: what one remembers of a remembered experience could not be enumerated in a closed

set of items of the sort needed to make out an isomorphism, and it would be impossible to devise a key of isomorphism that could provide any reasonable prospect for the discovery of a one-to-one correlation of component elements between any mental state and any neural state.

Philosophy of religion

Malcolm's paper "Anselm's Ontological Arguments" (1960) provoked considerable discussion. In it he says that Anselm put forward two different ontological proofs of the existence of God. The first, in Proslogion 2, uses the principle that a thing is greater if it exists than if it does not exist. The second, in Proslogion 3, uses the different principle that a thing is greater if it necessarily exists than if it does not necessarily exist. The first is fallacious because it is an error to regard existence as a property of things that have contingent existence, but it does not follow that it is an error to regard necessary existence as a property of God and as a perfection. A short summary of the second proof: If God exists, His existence is necessary; thus God's existence is either necessary or impossible; assuming that the concept of God is not self-contradictory or in some way logically absurd, it follows that He necessarily exists. Malcolm remarks, "I should think there is no more a presumption that [the concept of God] is self-contradictory than is the concept of seeing a material thing. Both concepts have a place in the thinking and the lives of human beings" (1963a: 160).

Bibliography of Malcolm's work

A complete list of Malcolm's articles published through 1981 may be found in Carl Ginet and Sydney Shoemaker (eds.) *Knowledge and Mind* (Oxford: Oxford University Press, 1983). The collection *Wittgensteinian Themes* (1995) contains fourteen of his essays written in the last twelve years of his life.

1942: "Certainty and Empirical Statements," *Mind* 51, pp. 18–46.
1949: "Defending Common Sense," *Philosophical Review* 58, pp. 201–21.
1950: "The Verification Argument," in *Philosophical Analysis*, ed. M. Black, Ithaca, NY: Cornell University Press. (Reprinted with revisions and additional footnotes in Malcolm 1963a.)
1951: "Philosophy for Philosophers" (intended title: "Philosophy and Ordinary Language"), *Philosophical Review* 60, pp. 329–40.
1956: "Dreaming and Skepticism," *Philosophical Review* 65, pp. 14–37.
1959: *Dreaming*, London: Routledge and Kegan Paul.
1960: "Anselm's Ontological Arguments," *Philosophical Review* 69, pp. 41–60. (Reprinted with new footnotes in Malcolm 1963a.)
1963a: *Knowledge and Certainty*, Englewood Cliffs, NJ: Prentice-Hall.
1963b: "Three Lectures on Memory," ("Memory and the Past," "Three Forms of Memory," and "A Definition of Factual Memory"), in Malcolm 1963a. ("Memory and the Past" first published in *The Monist* 45 (1962), pp. 247–66.)
1964: "Scientific Materialism and the Identity Theory," *Dialogue* 3, pp. 115–25.
1965: "Descartes' Proof that His Essence is Thinking," *Philosophical Review* 74, pp. 315–38. (Reprinted in Malcolm 1977b.)
1968: "The Conceivability of Mechanism," *Philosophical Review* 77, pp. 45–72.

1971: *Problems of Mind*, New York: Harper and Row.
1972: "Thoughtless Brutes," Presidential address, *Proceedings of the American Philosophical Association* 46, pp. 5–20. (Reprinted in Malcolm 1997b.)
1975: "Descartes' Proof that He is Essentially a Non-material Thing," *Philosophy Forum* 14 (1975). (Reprinted in Malcolm 1977b.)
1976: "Moore and Wittgenstein on the Sense of 'I know'," in *Essays in Honour of G. H. von Wright,* in *Acta Philosophica Fennica* ed. Jaakko Hintikka, 28, 1–3, pp. 216–40.
1977a: *Memory and Mind*, Ithaca, NY: Cornell University Press.
1977b: *Thought and Knowledge*, Ithaca, NY: Cornell University Press.
1980: "Functionalism in Philosophy of Psychology," *Proceedings of the Aristotelian Society,* new series 80, pp. 211–29.
1984: *Consciousness and Causality: A Debate on the Nature of Mind with D. M. Armstrong*, Oxford: Blackwell Publishers.
1986: *Wittgenstein: Nothing is Hidden*, Oxford: Blackwell Publishers.
1994: *Wittgenstein: A Religious Point of View?*, ed. with a response by P. Winch, Ithaca, NY: Cornell University Press.
1995: *Wittgensteinian Themes: Essays 1978–1989*, ed. G. Henrik von Wright, Ithaca, NY: Cornell University Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

19

Wilfrid Sellars (1912–1989)

JAY F. ROSENBERG

Life and work

Had Wilfrid Stalker Sellars never written an original philosophical word, his accomplishments as an editor would likely be sufficient to earn him a place of honor in the history of postwar analytic philosophy. In 1950, he and Herbert Feigl co-founded *Philosophical Studies*, the first scholarly journal explicitly devoted to analytic philosophy, which they edited jointly until 1971 and Sellars then edited alone for three more years. A year earlier, Feigl and Sellars had already published a seminal anthology, *Readings in Philosophical Analysis*; *Readings in Ethical Theory*, co-edited by Sellars and John Hospers, followed three years later. The "philosophical analysis" represented in these volumes had been transplanted from its origins and early development at Cambridge and Oxford and enriched by generous cross-fertilization from the logical empiricism of an expatriate Vienna Circle, most notably by the work of Rudolf Carnap, and indigenous strains of pragmatism, critical realism, and evolutionary naturalism. From such seeds, the "analytic" style of philosophizing and its agenda of problems grew to dominate American academic philosophy, definitively changing its intellectual landscape. The "logico-linguistic turn" became the new methodological center of philosophical inquiry, and regional philosophies of logic, language, mind, and science first joined and then gradually began to supplant more broadly-conceived traditional metaphysical and epistemological studies, while normative ethical inquiries gave ground to issues in metaethics and moral psychology.

But Wilfrid Sellars in fact wrote many an original and important philosophical word, and so not only helped to stimulate the growth of postwar analytic philosophy, but also became one of its most distinguished and influential practitioners. His academic trajectory took him from studies at the University of Michigan and, as a Rhodes Scholar, at Oxford University, to faculty positions at the University of Iowa, the University of Minnesota, where he served as Chair during the mid-1950s, and Yale University, before he became University Professor of Philosophy and Research Professor of the Philosophy of Science at the University of Pittsburgh, a position which he held from 1963 until his death in 1989. His intellectual trajectory meanwhile carried him from an early period, during which he worked out his fundamental philosophical ideas in over two dozen dialectically-challenging essays, through a middle period characterized by the

development and exposition of a systematic philosophical vision of remarkable scope and depth, into a late period of consolidation, refinement, and deepening of mature theses and insights that were simultaneously coming to be more fully appreciated and explicitly appropriated by a new philosophical generation.

The critique of givenness

Sellars's revolutionary 1956 essay "Empiricism and the Philosophy of Mind," immediately acknowledged as a contemporary classic, marks the beginning of his exceptionally productive and influential middle period. (This can be treated, somewhat arbitrarily, as culminating in 1972 with the publication of his 1970 American Philosophical Association Eastern Division Presidential Address on the Kantian text, "this I or he or it (the thing) which thinks.") The central theme of "Empiricism and the Philosophy of Mind" is a thoroughgoing and general critique of what Sellars famously called the "myth of the given," a perennial and polymorphic philosophical motif manifested *inter alia* in the idea, characteristic of classical sense-datum theory, that empirical knowledge rests on a foundation of "immediate awarenesses" and on the assumption that the "privacy" of the mental and one's "privileged access" to one's own mental states are primitive features of experience, logically and epistemologically prior to all intersubjective concepts pertaining to inner episodes.

Sellars criticized sense-datist and other traditional epistemic foundationalisms for failing properly to distinguish non-conceptual states of sensing from conceptually structured perceptual takings. Perceiving always involves taking something sensorily present *to be* this or that, and so, as Kant recognized, has a judgmental form which mobilizes and applies, correctly or incorrectly, descriptive and classificatory concepts. Perception properly so-called is consequently a normative business, and the ability to engage in it requires more than reliable differential dispositions to respond to sensory stimuli. "The essential point," Sellars wrote, "is that in characterizing an episode or a state as that of *knowing*, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says" (1963a: 169). A perceptual judgment may consequently be *direct* or *immediate* in the sense of being a unmediated response to stimulation, i.e. not itself inferred from other judgments, but the *epistemic authority* of such judgments depends upon their being appropriately caught up in the intersubjective game of having and giving reasons, and so cannot be independent of their inferential relationships to other judgments. Sellars thus notoriously advocated a strong *epistemic internalism*, according to which "observational knowledge of any particular fact . . . presupposes that one knows general facts of the form *X is a reliable symptom of Y*" (1963a: 168).

Epistemic authority

Consistent with his strong internalism, Sellars interpreted even first-person epistemic authority with respect to the sensory aspects of one's own experience as built on and presupposing an intersubjective status for sensory concepts *per se*. Correlatively, he decisively rejected the idea that sensory consciousness supplies a form of empirical knowledge that (1) is immediate (i.e. non-inferential); (2) presupposes no knowledge of other

matters of fact, particular or general; and (3) constitutes the ultimate court of appeals for all factual claims (1963a: 164). Although a person can justifiably believe an empirical truth without having inferred it from other propositions, no empirical beliefs are self-justifying, self-warranting, or self-authenticating. Instead, Sellars argued, "to say that someone directly knows that-p is to say that his right to the conviction that-p essentially involves the fact that the [thought] that-p occurred to the knower in a specific way" (p. 188).

The epistemic authority of a non-inferential perceptual belief can be traced to the fact that, in the course of learning perceptual language, the believer has not only acquired propensities for the reliable use of the relevant concepts in perceptual situations but also has come to know what is involved in learning to use perceptual sentences reliably in perceptual contexts. Thus, assuming that he has mastered the use of the relevant words in suitable perceptual situations, a person who candidly and spontaneously thinks-out-loud "Lo! Here is a red apple" – Sellars's customary model of a perceptual taking – is justified in reasoning:

> I just thought-out-loud "Lo! Here is a red apple" (no countervailing conditions obtain); so, there is good reason to believe that there is a red apple in front of me. (1975d: 341–2)

This "trans-level" reasoning, as Sellars called it, does not have the original perceptual judgment as its conclusion, but is rather an inference from the character and context of the original non-inferential experience to the existence of a good reason for accepting it as veridical.

Central to Sellars's thoroughgoing epistemic internalism, indeed, is his conviction that the reasonableness of accepting even *first principles* is a matter of the availability of good arguments warranting their acceptance. What is definitive of first principles (FP) is the unavailability of sound arguments in which they are derived as conclusions from still-more-basic premises, that is, arguments of the form:

(A1) $P_1, P_2, \ldots, P_n \vdash$ FP.

Here, too, Sellars appeals to the notion of a "trans-level" justificatory inference, pointing out that the absence of good arguments of the form (A1) is entirely compatible with the existence of sound arguments of the form

(A2) $P_1, P_2, \ldots, P_m \vdash$ It is reasonable to accept FP.

Sellars interpreted the conclusion of (A2) as a claim to the effect that a particular course of epistemic *conduct*, accepting the principle FP, can be supported by adequate reasons. This suggested the in-principle availability of yet another argument, specifically a piece of sound *practical* reasoning, whose conclusion expresses an intention to engage in such conduct:

(A3) I shall achieve desirable epistemic end E.
Achieving E implies accepting principles of kind K.
The principle FP is of kind K.
Therefore, I shall accept FP.

On Sellars's view, such patterns of practical reasoning also govern the warranted acceptance of lawlike generalizations and theoretical systems. Adopting a systematic

theoretical framework is ultimately justified by appeal to the epistemic end of "being able to give non-trivial explanatory accounts of established laws" (1975c: 384), and adopting nomological claims that project the observed statistical frequency (including a frequency = 1) of some property in an open class to further unobserved finite samples of the class, is ultimately justified by appeal to the epistemic end of "being able to draw inferences concerning the composition with respect to a given property Y of unexamined finite samples . . . of a kind X, in a way which also provides an explanatory account of the composition with respect to Y of the total examined sample, K, of X" (1975c: 392). What is crucial is that these ends concern

> the realizing of a logically necessary condition of being in the framework of explanation and prediction, i.e. being able to draw inferences concerning the unknown and give explanatory accounts of the known. (1975c: 397)

Sellars argued, in short, that inductive reasoning does not need to be vindicated, that is, shown to be truth-preserving, but is rather itself fundamentally a form of vindicative reasoning, justifying our engaging in determinate epistemic conducts. Its ends-in-view must consequently be of a sort that can be known to be realized or obtain. Unlike such Reichenbachian ends as being in possession of limit-frequency statements which are within a certain degree of approximation of the truth (where such limits exist), the ends of being in possession of explanatory laws and principles envisioned by Sellars satisfed that constraint.

Self-knowledge

Sellars famously engaged the Cartesian picture of direct and incorrigible *self*-knowledge with his "myth of Jones," a story set in a fictional community whose members speak a hypothetical sophisticated "Rylean language." The fundamental descriptive vocabulary of this language pertains to public spatiotemporal objects, and while it includes logical operators, subjunctive conditionals, and even the fundamental resources of semantic discourse (enabling its speakers to say of their peers' utterances that they mean this or that, stand in various logical relations to one another, and are true or false), it nevertheless lacks any resources for speaking of inner episodes, whether thoughts or experiences.

> In this community, then, a genius, Jones, develops a theory according to which overt utterances are but the culmination of a process which begins with certain inner episodes. . . . [His] model for these episodes which initiate the events which culminate in overt verbal behavior is that of overt verbal behavior itself. (1963a: 186)

Jones's theory earns its epistemic credentials by enabling him, and his fellow Ryleans who master it, successfully to explain and anticipate intelligent behavior conducted in silence, that is, unaccompanied by overt verbal episodes of the sort that we would recognize as expressing an agent's conduct-rationalizing beliefs and desires.

The new idioms of Jones's theory, for example, "is thinking '. . .'," initially have a purely theoretical use, being ascribed on the basis of inferences from observable

behavior in observable circumstances. But crucially, Sellars argued, what begin as purely theoretical terms can *acquire* a first-person reporting role. For it turns out to be possible for Jones to train his compatriots, in essence by a process of Skinnerian operant conditioning, to respond reliably and directly (i.e. non-inferentially) to the occurrence of such an "inner episode" with a judgment to the effect that it is occurring. That is, they can respond to one thought with a second (meta-)thought to the effect that they are thinking it; this matches the *de facto* phenomenology of first-person "privileged access" sufficiently to account for the Cartesian illusion of mental transparency. The Jonesean story thus shows how the essential intersubjectivity of language can be reconciled with the "privacy" of inner episodes.

Scientific realism

Sellars's novel appeal to forms of theoretical reasoning in his myth of Jones reflected his broader philosophical concern with the nature, structure, and role of theories in the natural sciences. On the received, positivist, view explanation was identified with derivation. Singular matters of empirical fact were to be explained by deriving descriptions of them from ("inductive") empirical generalizations (along with appropriate statements of initial conditions), and these "empirical laws" in turn were to be explained by deriving them from theoretical postulates and correspondence rules. On the positivist view, in consequence, theories (e.g. microtheories) explain observational matters of fact only indirectly, by implying the (observation-language) generalizations that explain them directly.

Sellars, in contrast, argued that this "levels picture" of theories was fundamentally misleading. Theories do not explain laws by entailing them. Rather, "theories explain laws by explaining why the objects of the domain in question obey the laws that they do to the extent that they do" (1963c: 123).

> [That is,] they explain why individual objects of various kinds and in various circumstances in the observation framework behave in those ways in which it has been inductively established that they do behave. Roughly, it is because a gas is . . . a cloud of molecules which are behaving in certain theoretically defined ways, that it obeys the empirical Boyle–Charles Law. (1963c: 121)

On Sellars's view, then, "theoretical entities" are not merely convenient fictions, enabling us to abbreviate complicated and unwieldy stories about entities that we have good, (observational) reasons to believe actually exist. Theoretical entities are rather those entities we justifiedly believe to exist for good and sufficient *theoretical reasons*.

Sellars thus advocated a robustly realistic epistemology of scientific inquiry and, correlatively, an understanding of its ultimate outcomes as ontologically definitive: "In the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not" (1963a: 173).

(This is his "scientia mensura.") Scientific theories indeed explanatorily "save the appearances," but they do so precisely by describing the reality of which the appearances are appearances. This robust realism, combined with a thoroughgoing

naturalism, in fact set Sellars's metaphilosophical agenda for postwar analytic philosophy *per se*.

Metaphilosophical views

Sellars saw contemporary philosophy as confronted by two "images," each of which purported to be a complete picture of man-in-the-world, which need to be unified into a single synoptic vision. The "manifest image" is the conception of the world and the place of persons in it that has descended from the great speculative systems of ancient Greece, through the dialectics of a "perennial philosophy," to the dimensions of contemporary Anglo-American thought that emphasize "ordinary usage" and "common sense." Its primary objects are persons, beings who, *inter alia*, reflectively conceive of themselves as being in the world both as sentient perceivers and cognitive knowers of it, and as agents capable of affecting it through deliberate and rational elective conduct.

In contrast, the "scientific image" is the complex new understanding of man-in-the-world that is still in the process of emerging from the fruits of theoretical reasoning, in particular, from the processes of postulational theory construction. Although this image is "methodologically dependent on the world of sophisticated common sense," Sellars argues that

> it purports to be a complete image, i.e. to define a framework which could be the whole truth about that which belongs to the image. Thus although methodologically a development within the manifest image, the scientific image presents itself as a rival image. From its point of view the manifest image on which it rests is an "inadequate" but pragmatically useful likeness of a reality which first finds its adequate (in principle) likeness in the scientific image. (1963c: 57)

The leading challenge for contemporary philosophy, he concluded, is to show how the inevitable tensions between these two images can be resolved by a "stereoscopic understanding" in which they come to be "fused" into a single synoptic vision of man-in-the-world. Sellars's philosophy is usefully viewed as a fuller articulation of this confrontation of the two images and a detailed dialectical engagement with the philosophical agenda to which it gives rise: that places be found within the context of a thoroughgoing naturalism for the intentional contents of language and thought, for the normative dimensions of knowledge and action, and for the sensuous contents of perception and imagination. Consonant with such a naturalism, the sought synoptic story must find a place for mind without assigning an independent, autonomous, and irreducible, ontological status to intentional states or entities, and it must eschew any ontological view of abstracta as real that fails to deliver an adequate account of their role within the causal order, broadly construed.

Semantic meaning

The centerpiece of Sellars's response to both of these naturalistic challenges was a sophisticated theory of conceptual roles, concretely instantiated in the conducts of linguistic communities and socially transmitted by modes of cultural inheritance. At the

heart of this theory was an increasingly refined account of meaning as functional classification, more precisely, of the "meaning" idiom as, in the first instance, a context of translation in terms of which structurally distinct "natural-linguistic objects" (e.g. utterings or inscribings) are classified in terms of their roles or functions *vis-à-vis* the organized behavioral economies of families of speaking organisms. In short, Sellars interpreted "means" as a specialized form of the copula, tailored to metalinguistic contexts, according to which the right side of the superficially relational form "– means . . ." is also properly understood as mentioning or exhibiting a linguistic item.

Ordinary quotation, argued Sellars, suffers from a systematic ambiguity regarding the criteria according to which linguistic tokens are correctly classifiable as belonging to this or that linguistic type. He therefore introduced the straightforward device of two separate styles of quotation marks – star-quotes and dot-quotes – to differentiate between two different ways of sorting and individuating lexical items. Star-quotes form common nouns that are true of items belonging to the spatiotemporal causal order ("tokens") which are appropriately *structurally* isomorphic to the tokens exhibited between them, while dot-quotes form common nouns true of tokens that, in some specified language, are appropriately *functionally* isomorphic to (i.e. play the role or do the job performed by) the tokens exhibited between them in *our* language. Sellars then proposed to transcribe such semantic claims as

(1s) (In French) "rouge" means *red*,
and
(2s) (In German) "Es regnet" means *it is raining*,

by the more perspicuous formulations

(1*) (In the French linguistic community) *rouge*s are ·red·s
and
(2*) (In the German linguistic community) *Es regnet*s are ·it is raining·s.

Roles and rules

To sort and classify descriptively individuated families of natural items in terms of their linguistic jobs, roles, or functions, is to sort and classify them *normatively*. Sellars's overall philosophical agenda thus required a complementary thoroughly naturalistic account of the normative dimension of language, and he indeed offered one, basing it on the notion of what he called *pattern-governed behavior*. The basic concept of pattern-governed behavior is

> the concept of behavior which exhibits a pattern, not because it is brought about by the intention that it exhibit this pattern, but because the propensity to emit behavior of the pattern has been selectively reinforced, and the propensity to emit behavior which does not conform to this pattern selectively extinguished. (1974: 423)

Pattern-governed behavior can arise from processes of natural selection on an evolutionary time-scale as a characteristic of a species, for example, the dance of the bees,

but it can also be developed in individuals, "trainees," by deliberate and purposive selection on the part of other individuals, the "trainers." In this connection, Sellars introduced a distinction between "rules of action" and "rules of criticism."

Rules of action specify what someone ought to do, for example, "*Ceteris paribus*, one ought to say such and such if in circumstances C." They can be efficacious in guiding *linguistic* activity only to the extent that their subjects already possess the relevant concepts, such as concepts of "saying such-and-such," of "being in circumstances C," and, indeed, of obeying a rule (i.e. doing something because it is enjoined by a rule). Rules of criticism, in contrast, specify what ought to be the case, for example, "Westminster clock chimes ought to strike on the quarter hour" (1975d: 95). The items whose performances may legitimately be appraised according to such rules, however, need not themselves have the concept of a rule nor, indeed, even be capable of having any concepts at all.

Thus trainers can be understood in the first instance as acting in accordance with rules of conduct whose authority derives from rules of criticism, that is, as aiming at bringing about the pattern-governed behaviors which their trainees' conduct ought to manifest: "Patterned-behavior of such and such a kind ought to be exhibited by trainees, hence we, the trainers, ought to do this and that, as likely to bring it about that it is exhibited" (1974: 423).

If training is successful, then, in consequence of the conducts of trainers under the guidance of such rules of action, the behavior of a language-learner can come to conform to the relevant rules of criticism without his, in any other sense, grasping them himself.

> [The] members of a linguistic community are first language learners and only potentially "people," but subsequently language teachers possessed of the rich conceptual framework this implies. They start out being the subject matter of the ought-to-be's and graduate to the status of agent subjects of the ought-to-do's. (1975d: 100)

The modes of pattern-governed behavior relevant to semantic meaning and, correspondingly, the relevant families of rules of criticism, Sellars proposed, are threefold:

1 Language Entry Transitions: It ought to be the case that speakers respond to objects in perceptual situations and to certain states of themselves with appropriate linguistic-conceptual activity.
2 Intra-linguistic Moves: It ought to be the case that speakers' linguistic-conceptual episodes tend to occur in patterns of valid inference (theoretical and practical, formal and material), and tend not to occur in patterns which violate logical principles.
3 Language Departure Transitions: It ought to be the case that speakers respond to such first-person linguistic-conceptual episodes as "I shall now raise my hand" with an upward motion of the hand, etc. (Cf. 1974: 423–4)

These transitions are respectively the core elements of perceptual takings, inferences, and volitions, and, although they are acts (*qua* both actualities and actualizations of behavioral dispositions), Sellars insisted that they are not themselves actions. They are acquired as, and remain, pattern-governed activities, but form the basis of linguistic

actions proper as "the trainee acquires not only the repertoire of pattern-governed linguistic behavior which is language about non-linguistic items, but also that extended repertoire which is language about linguistic as well as non-linguistic items" and thus becomes "able to classify items in linguistic kinds, and to engage in theoretical and practical reasoning about his linguistic behavior" (1974: 425).

Linguistic roles or functions are individuated in terms of the structure of positive and negative uniformities generated in the natural order by these pattern-governed activities of perception, inference, and volition. Analyzing sameness of linguistic role as sameness of place in the complex relational structure generated by conducts that have been causally shaped in these ways by systems of espoused linguistic norms equipped Sellars with a functional conception of semantics. This conception neither presupposed nor unavoidably led back into the domains of either ontological abstracta or irreducibly intentional mental entities.

The intentionality of thought

Instead, Sellars's alternative account of the distinctive intentionality of thought was itself drawn in terms of the forms and functions of natural linguistic items, modeled by what he came to call "verbal behaviorism" (VB):

> According to VB, thinking "that-p," where this means "having the thought occur to one that-p," has as its primary sense [an event of] saying "p"; and a secondary sense in which it stands for a short term proximate propensity [disposition] to say "p." (1974: 419)

The roots of Sellars's mature verbal behaviorism reach back to the myth of Jones in "Empiricism and the Philosophy of Mind." The thought-episodes postulated by Jones on the model of overt verbal behavior are introduced by a purely functional analogy. The concept of an occurrent thought is not that of something encountered *propria persona* but rather that of a causally-mediating logico-semantic role-player, whose determinate ontological character is initially left entirely open.

Since on Sellars's account the concept of a thought is fundamentally the concept of a functional kind, no ontological tension is generated by the identification of items belonging to that functional kind with states and episodes of an organism's central nervous system. The manifest image's conception of persons as thinkers can consequently fuse smoothly with the scientific image's conception of persons as complex material organisms having a determinate physiological and neurological structure. Sellars's conviction that the fundamental characteristic of semantic discourse is its ineliminable appeal to functional considerations, and his correlative pioneering analyses of the intentional categories of the mental in terms of epistemologically theoretical transpositions of the semantic categories of public language earn him the title of the first *functionalist* in contemporary analytic philosophy of mind. This functionalist view is one whose implications and influence have not yet begun to be exhausted.

Categorial ontology

The parallels between semantic discourse and the classical ontological idioms of Platonistic discourse, ostensibly designating abstract entities, have not gone unnoticed.

Consistent with his global commitment to naturalism, Sellars exploited these parallels to construct his own unique variant of linguistic nominalism, a view according to which

> the abstract entities which are the subject matter of the contemporary debate between platonistic and anti-platonistic philosophers – qualities, relations, classes, propositions, and the like – are linguistic entities. (1967a: 229)

In first approximation, Sellars proposed to analyze the ostensibly relational ontological claims

(1o) (The French word) "rouge" stands for *redness*,

and

(2o) (The German sentence) "Es regnet" expresses the proposition *that it is raining*,

precisely as he had analyzed the corresponding semantic claims (1s) and (2s). These two will be analyzed as (1*) and (2*), claims that specify the functional roles of determinate families of structurally-individuated tokens. This strategy of understanding traditional ontological discourse as classificatory discourse within a functional metalanguage transposed into the "material mode of speech" had been pioneered by Carnap (see CARNAP). But unlike Carnap, Sellars refused to identify the formally definable constructs of a "pure" syntax or semantics with the corresponding syntactical and semantical terms in everyday, pre-philosophical usage. Such a facile interpretation of the relationship between "pure" and "descriptive" syntactic and semantic discourses, he argued, fails to do proper justice to the essential normative dimension of the latter. Thus, while Sellars was prepared to interpret such paradigmatic ontological terms as "universal," "individual," "kind," "quality," "proposition," and "fact" by appealing to syntactic and semantic counterparts (e.g. "predicate," "singular term," "common noun," "monadic predicate," "sentence," and "true sentence") he insisted that these syntactic and semantic terms

> have a conceptual role which is no more reducible to [non-syntactical and] non-semantical roles than the role of prescriptive terms is reducible to non-prescriptive roles. . . . [The] empirical (in the broad sense) character of statements in descriptive (historical) [syntax and] semantics does not entail that [syntactical and] semantical concepts, properly so called, are descriptive. (1975a: 459)

Like Sellars's account of the distinctive intentionality of thought, then, his account of discourse ostensibly about the entities and categories of classical ontology is also drawn in terms of the forms and functions of natural linguistic items. "Abstract entities," too, consequently constitute no obstacle to the sought fusion of the manifest and scientific images.

Sensations

Surprisingly it is when Sellars turns from intentional thought and ontological abstracta to *sensations* that significant complications to his synoptic project first come into view.

Like Kant, Sellars rejected the Cartesian picture of a sensory-cognitive continuum. The "of-ness" of a sensation – e.g. its being of a red triangle or of a sharp shooting pain – he insisted, is not the intentional "of-ness" ("aboutness") of a thought. "The 'rawness' of 'raw feels'," he wrote, "is their non-conceptual character" (1967c: 376). Consequently, although Sellars's *epistemological* story about sensations also begins with a strategic appeal to the unique epistemic status of postulated theoretical entities, his account of the *ontology* of sensations diverges significantly from his semantic and functionalist account of intentional thoughts.

> In a final episode of the Myth of Jones the hero . . . postulates a class of inner – theoretical – episodes which he calls, say, impressions, and which are the end results of the impingement of physical objects and processes on various parts of the body. (1963a: 191)

Jones postulates impressions as elements of an explanatory account of the occurrence in various circumstances of perceptual cognitions with determinate semantic contents. In this case, however, the *model* for Jones's theory is not functionally-individuated families of sentence tokens, but rather "a domain of 'inner replicas' which, when brought about in standard conditions share the perceptible characteristics of their physical sources" (1963a: 191). Although the entities of this *model* are particulars, the entities introduced by Jones's *theory* are not particulars but rather non-conceptual (non-intentional) states of a perceiving subject. Thus, while talk of the "of-ness" of sensations, like that of the "of-ness" of thoughts, is fundamentally classificatory, the classification of sensations is not functional, but rather based on analogies that are initially extrinsic and causal, and ultimately intrinsic and contentive.

In the first instance, then, the concept of a person's having an of-a-red-triangle sensation – an adjectival idiom contrived to highlight the classificatory role of "of-ness" – or the concept of a person's sensing $[\text{red triangle}]_S$ly – an adverbial idiom contrived to reflect the fact that "sensation" is a "verbal noun" – is the concept of her being in the sort of state that is brought about in normal perceivers in standard conditions by the action of red triangular objects on the eyes. The point of the model of "inner replicas," however, is to insist that such states can discharge their explanatory jobs in relation to cognitive perceptual takings (especially non-veridical perceptual judgments) only if they are conceived as having themselves determinate intrinsic characters and, in particular, as resembling and differing from other sensory states in a manner formally analogous to the way in which objects of the "replica" model (e.g. "wafers" of various colors and shapes) are conceived to resemble and differ from one another.

Sellars proceeded to develop this core account of sensations in two different directions, in consequence of which it has come to be regarded as one of the most difficult and controversial aspects of his philosophy. The first line of development turned on his conclusion that the fundamental concept pertaining to color within the manifest image is the concept of a kind of stuff. It is the concept of a quantum of red in space, an expanse or volume consisting of red, and is a *basic* concept in the sense that there is "no . . . determinate category prior to the concept of red as a physical stuff, as a matter for individuated physical things" (1981, I: 84). When dialectical pressures generate worries about the ontological status of the red which one ostensibly sees when it is *not* a constituent redness of a physical object, however, no alternative category can simply

be "read off" from an introspective scrutiny of color quanta. The idea that a person is always also aware of the actual categorial status of the items that he encounters in perception or introspection, Sellars suggests, is only a particularly pernicious form of the myth of the given.

> All that is available is such transcendentals as "actual," "something" and "somehow." The red is something actual which is somehow a portion of red stuff, somehow the sort of item which is suited to be part of the content of a physical object, but which . . . is not, in point of fact, a portion of physical stuff. (1981, I: 90)

It then becomes the job of analogical thinking to construct new categorial forms of concepts pertaining to color.

The first complication of Sellars's account of sensation resulted from his conviction that, in the case of sensations, Jones's theory takes this *interpretive* form. It does not introduce new domains of entities, but rather new forms of concepts.

> [Jones's] theory of sense impressions . . . reinterprets the categorial status of the cubical volumes of pink of which we are perceptually aware. Conceived in the manifest image as, in standard cases, constituents of physical objects and in abnormal cases, as somehow "unreal" or "illusory," they are recategorized as sensory states of the perceiver and assigned various explanatory roles in the theory of perception. (1981, III: 44)

The crux of the Jonesean theory, in other words, is the thesis that the very color quanta of which we are perceptually aware as being in space are instead actually states of persons-qua-perceivers. It follows that, already within the manifest image, the ontological status ultimately accorded to sensory "content qualia" is in fact incompatible with their actually being instantiated in physical space. "[The] *esse* of cubes of pink is *percipi* or, to use a less ambiguous term, *sentiri*. Of course . . . we are not perceptually aware of [them] as states of ourselves, though that is in point of fact what they are" (1981, III: 66).

Absolute processes

The second complication of Sellars's account of sensations then arose from his further conclusion that the scientific image's commitment to the idea that perceivers are complex systems of micro-physical particles constitutes a barrier to any straightforward synoptic assimilation of this manifest image conception of sensory contents as states of perceiving subjects. On the one hand, Sellars observed that Jones's analogical treatment of sensory contents as states of perceivers formally preserves the "ultimate homogeneity" of those contents as originally conceived as space-filling stuffs. No defined states of a system or multiplicity of logical subjects, he argued, could continue to do so.

On the other hand, Sellars contended, we cannot simply adopt a "reductive materialist" view according to which "what really goes on when a person senses a-cube-of-pinkly consists in [a certain] system of micro-physical particles being in a complex physical-2 state" (1981, III: 79), where "physical-2" states are those definable

in terms of theoretical predicates necessary and sufficient to describe non-living matter. (To be "physical-1," in contrast, is simply to belong in the space-time network.) For such reductive materialism amounts to the rejection of the idea that a (Jonesean-theoretical) state of, for example, sensing a-cube-of-pinkly is itself something actual in any categorial guise, and this fails properly to respect the philosophical demands of an adequate sensory phenomenology.

What Sellars notoriously concluded was that sensory contents could be synoptically integrated into the scientific image only if *both* they *and* the micro-physical particulars of that image as well were subjected to yet another ontological reconception in terms of a categorially-monistic framework whose basic entities were all "absolute processes." Only when perceivers themselves had been reconceived as systems or "harmonies" of absolute processes, including the ultimate conceptual descendants of sensory contents, would the way be cleared for a unification of the two images. Thought of as absolute processes, sensings would be physical

> not only in the weak sense of not being mental (i.e. conceptual), for they lack intentionality, but in the richer sense of playing a genuine causal role in the behavior of sentient organisms. They would . . . be physical-1 but not physical-2. Not being epiphenomenal, they would conform to a basic metaphysical intuition: to be is to make a difference. (1981, III: 126)

Intention and action

In contrast to the integrative challenges posed by thoughts and sensations, the challenge of integrating *actions* properly so called, that is, conducts informed by intention and volition, into the scientific image is not fundamentally ontological. Although they exhibit quite special features when considered functionally, regarded from the ontological perspective, intentions and volitions are simply species of occurrent thought-episodes. What makes such thoughts *practical* are their special relationships to conduct or behavior, analogous to the way in which their status as non-inferential responses to sensations confers on particular thoughts the functional role of perceptual judgments.

Sellars characteristically signals the special conduct-determining role of practical thoughts by a contrived use of the auxiliary verb "shall" as an operator on thought contents expressed as sentences. Categorical intentions are temporally determinate first-person future-tensed practical thoughts. They have the canonical form (INT): Shall (I will do X at t). Volitions ("acts of will") are special cases of such intentions, whose time determination is the indexical present, thus, (VOL): Shall (I will *now* do X).

Such practical thinkings, on Sellars's view, mediate between deliberative reasoning and overt behavior by being appropriately caught up in a network of acquired causal propensities that guarantee, roughly, that intentions of the form (INT) regularly give rise, at time t, to volitions of the form (VOL), which, absent paralysis and the like, in turn regularly give rise, then and there, to bodily movements that are (further circumstances being appropriate) the initial stages of a doing of X. Such practical thinkings are governed according to a single principle which unites practical and theoretical reasoning: If $\ulcorner p \urcorner$ implies $\ulcorner q \urcorner$, then $\ulcorner \text{Shall}(p) \urcorner$ implies $\ulcorner \text{Shall}(q) \urcorner$.

Persons

Here again, then, Sellars concluded, what the manifest image contains is not the concept of something with a determinate intrinsic character with which we are acquainted but rather the functional conception of a causally-mediating logico-semantic role-player. Practical thinkings can thus be ontologically accommodated within the scientific image along the lines already sketched for cognitive thoughts in general. But here, he continued, ontological accommodation cannot be the end of the synoptic story. If we take seriously the idea that the scientific image purports to be a complete image of man-in-the-world and a candidate ultimately to replace the manifest image, then the latter's categories pertaining to *persons* will need to reappear within the sought synoptic fusion as such. We need to reconcile "the idea that man is what science says he is" with "the categories pertaining to man as a person who finds himself confronted by standards (ethical, logical, etc.) which often conflict with his desires and impulses" (1963c: 38).

On Sellars's view, the basic concept of a person is irredeemably social. To think of an entity as a person is essentially to think of it as actually or potentially a member of a community, "an embracing group each member of which thinks of itself as a member of the group" (1963c: 39), and it is the most general shared intentions of its members that fundamentally define the structure of norms and values in terms of which their conducts come to be appraised as "correct" or "incorrect" or "right" or "wrong."

> Thus the conceptual framework of persons is the framework in which we think of one another as sharing the community intentions which provide the ambience of principles and standards (above all, those which make meaningful discourse and rationality itself possible) within which we live our own individual lives. (1963c: 39–40)

As we have seen, Sellars interpreted the framework of thoughts as founded within the manifest image on a series of ontologically noncommittal functional analogies to which we can readily imagine an emerging scientific understanding progressively supplying structural (e.g. neurophysiological) form. In contrast, he argued that accommodating the manifest image's sensory contents within a synoptic fusion would require the conceptual transposition of some of its ontologically basic entities into new categorial forms. Unlike the frameworks of thoughts and sensations, however, Sellars contended that the conceptual framework of persons as such "is not something that needs to be reconciled with the scientific image, but rather something to be joined to it" (1963c: 40). To achieve a genuinely synoptic vision of man-in-the-world, he concluded, the scientific image needs to be enriched

> not with more [or different] ways of saying what is the case, but with the language of community and individual intentions, so that by construing the actions we intend to do and the circumstances in which we intend to do them in scientific terms, we directly relate the world as conceived by scientific theory to our purposes, and make it our world and no longer an alien appendage to the world in which we do our living. (1963c: 40)

Such an ultimate unification of the manifest and scientific images, the world of persons with the world of science, was the controlling vision of Sellars's philosophy.

What makes him one of the towering figures of postwar analytic philosophy, however, is not just the grand scope of his enterprise, but the profound originality of his specific conclusions, the sophisticated dialectically argued and historically informed reasoning with which he supported them, and the exemplary thoroughness with which he painstakingly developed something very rare in the analytic tradition, a principled and consistent *systematic* philosophical view.

Bibliography of works by Sellars

1963a: "Empiricism and the Philosophy of Mind," in 1963d, pp. 127–96. (Formerly published in *The Foundations of Science and the Concepts of Psychoanalysis*, Minnesota Studies in the Philosophy of Science, vol. I, ed. H. Feigl and M. Scriven, Minneapolis: University of Minnesota Press, 1956.)

1963b: "Phenomenalism," in 1963d, pp. 60–105.

1963c: "Philosophy and the Scientific Image of Man," in 1963d, pp. 1–40. (Formerly published in *Frontiers of Science and Philosophy*, ed. Robert Colodny, Pittsburgh, PA: University of Pittsburgh Press, 1962.)

1963d: *Science, Perception and Reality*, London: Routledge and Kegan Paul, and New York: Humanities Press. (Reissued by Ridgeview Publishing.)

1963e: "The Language of Theories," in 1963d, pp. 106–26. (Formerly published in *Current Issues in the Philosophy of Science*, ed. H. Feigl and G. Maxwell, New York: Holt, Rhinehart and Winston, 1961.)

1967a: "Abstract Entities," in 1967b, pp. 229–69. Reprinted in *Review of Metaphysics* 16 (1983).

1967b: *Philosophical Perspectives*, Springfield, IL: Charles C. Thomas. (Reissued in 2 vols, *Philosophical Perspectives: History of Philosophy* and *Philosophical Perspectives: Metaphysics and Epistemology*, by Ridgeview Publishing.)

1967c: "The Identity Approach to the Mind–Body Problem," in 1967b, pp. 370–88. (Formerly published in *Review of Metaphysics* 18 (1965).)

1968: *Science and Metaphysics: Variations on Kantian Themes*, London: Routledge and Kegan Paul and New York: Humanities Press. (Reissued by Ridgeview Publishing.)

1974: "Meaning as Functional Classification," *Synthese* 27, pp. 417–37.

1975a: "Empiricism and Abstract Entities", in 1975b, pp. 245–86. (Formerly published in *The Philosophy of Rudolph Carnap*, ed. P. A. Schilpp, La Salle, IL: Open Court, 1963.)

1975b: *Essays in Philosophy and its History*, Dordrecht: Reidel.

1975c: "Induction as Vindication," in 1975b, pp. 367–416. (Formerly published in *Philosophy of Science* 31 (1964).)

1975d: "The Structure of Knowledge," in *Action, Knowledge, and Reality: Studies in Honor of Wilfrid Sellars*, ed. H. N. Castaneda, Indianapolis: Bobbs-Merrill, pp. 295–347.

1975e: "Language as Thought and Communication," in 1975b, pp. 93–117. (Formerly published in *Philosophy and Phenomenological Research* 29 (1969).)

1979: *Naturalism and Ontology*, Reseda, CA: Ridgeview Publishing.

1980: *Pure Pragmatics and Possible Worlds: The Early Essays of Wilfrid Sellars*, ed. J. F. Sicha, Reseda, CA: Ridgeview Publishing. (Also contains a long introductory essay by Sicha and an extensive bibliography of Sellars's work through 1979.)

1981: "The Carus Lectures for 1977–78," *The Monist* 64/1. (Citations by lecture and numbered paragraph.)

1989: *The Metaphysics of Epistemology, Lectures by Wilfrid Sellars*, ed. P. Amaral, Reseda, CA: Ridgeview Publishing. (Also contains a complete bibliography of Sellars's published work through 1989.)

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

20

H. P. Grice (1913–1988)

STEPHEN NEALE

Life

Herbert Paul Grice was born on March 13, 1913, in Birmingham, England. He attended Corpus Christi College, Oxford, graduating in 1936. From 1938 until 1967 he held various fellowships and lectureships at St John's College. His time at Oxford was interrupted by nearly five years' wartime service in the Royal Navy, first in the North Atlantic and later in Admiralty intelligence. In 1967, he moved to the University of California, Berkeley as Professor of Philosophy. He was elected to the British Academy in 1966, and gave the William James Lectures at Harvard in 1967, the John Locke lectures at Oxford in 1978, and the Tanner Lectures at Stanford in 1980. He died in Berkeley in August 1988, shortly before the publication of his first book, *Studies in the Way of Words.*

Grice was one of the most gifted and respected philosophers of the second half of the twentieth century. He set impossibly high standards and was always reluctant to go into print – heroic efforts were required by editors and friends to extract from him the handful of papers he deemed worthy of publication – yet he exerted considerable influence through seminars and invited lectures. He worked on topics in Aristotle, metaphysics, ethics, and philosophical psychology; but his strongest influence was in the philosophy of language, where his thought continues to shape the way philosophers, linguists, and cognitive scientists think about meaning, communication, and the relation between language and mind.

With respect to a particular sentence, *X*, and an "utterer" *U*, Grice stressed the importance of separating (1) what *X means*, (2) what *U said* on a given occasion by uttering *X*, and (3) what *U meant* by uttering *X* on that occasion. Second, he attempted to say what meaning *is* by providing analyses of utterer's meaning, sentence meaning, and what is said. Third, he tried to explain how what *U* says and what *U* means can diverge. Fourth, he defended conceptual analysis and some form of analytic/synthetic distinction. Fifth, by characterizing the distinction between the "genuinely semantic" and "merely pragmatic" implications of a statement, Grice clarified the relationship between classical logic and the semantics of natural language. Sixth, he deployed his notion of "implicature" to devastating effect against overzealous strains of "ordinary-language philosophy," without abandoning the view that philosophy must pay atten-

tion to the nuances of ordinary talk (see AUSTIN). Seventh, Grice undercut the most influential arguments for a philosophically significant notion of "presupposition." Eighth, he made significant contributions to debates about the semantics of proper names, definite descriptions, and pronouns. Ninth, he sketched a philosophical psychology and a theory of value that promise to provide the basis of future work on actions, mental states, and moral philosophy, and to explain the relationship between mind and language inherent in his philosophy of language.

Meaning, use, and ordinary language

The view that the only useful thing to say about the meaning of an expression is that it is usable in such-and-such circumstances, exercised a powerful influence on philosophy in postwar Oxford. Austin, Ryle, and others undercut philosophical positions or disposed of philosophical problems by pointing to a misuse of some expression playing an essential role in the presentation of the position or problem. Consider attempts to analyze *knowledge* in terms of *belief* along the following lines: *A* knows that *p* if and only if (1) *A* believes that *p*, (2) *p*, and (3) *A* is justified in believing that *p*. It might be charged that it is a feature of the use of "believe" that one does not use it if one can sincerely use "know" instead. Such a claim might be supported by observing that it would be inappropriate for a man to say "I believe Smith is dead" when he knows Smith is dead. And so it might be concluded that the proposed analysis must be discarded because clause (1) conflicts with the ordinary use of the verb "believe."

Grice accepted that a theory of meaning must be sensitive to use and attempted to explicate the meaning of an expression (or any other sign) in terms of what its users *do with it*, that is, in terms of what its users (could/would/should) mean by it on particular occasions of use. Two important ideas came out of this sensitivity to use. The first is that the locution *by uttering x, U meant that p* can be analyzed in terms of complex audience-directed intentions on the part of *U*. The second is that the most "basic" notion of meaning is that of an utterer *U* meaning something by doing something on a particular occasion; all other notions of meaning are derivative. What *U* means by producing *x* on a given occasion is a function of what *U* intends, in a complex way, to *get across* to his audience. The basic idea is, very roughly, that for an "indicative-type" utterance, the locution *by uttering x, U meant that p* expresses a truth iff *U* uttered *x* intending to produce in some audience *A* the belief that *p* by means of *A*'s recognition of this intention. Sentence meaning is to be analyzed in terms of regularities over the intentions with which utterers produce sentences on given occasions.

By uttering a sentence of the form "*p* or *q*," *U* may well imply that he has non-truth-functional grounds for his assertion; but this is not part of what the sentence (or the statement made) implies. Grice wanted any adequate explanation of the possibility of pragmatic implications to flow from a completely *general* theory. To demonstrate the definite existence of pragmatic implications distinct from semantic implications, Grice considered an extreme example. Suppose *A* asks *U* for an evaluation of his student Mr. *X*. All *U* says is "Mr. *X* has excellent handwriting and is always very punctual." If *U* leaves it at that, those present are likely to conclude that *U* thinks Mr. *X* is not much good at philosophy. There is surely no temptation to say that the proposition that Mr. *X* is not much good (or that *U* thinks Mr. *X* is not much good) at philosophy is (or is a conse-

quence of) the statement *U* made. The sentence *U* uttered has a clear linguistic meaning based on the meanings of its parts and their syntactical arrangement; and it seems quite wrong to say that, when he uttered that sentence, *U* made the statement that Mr. *X* is not much good at philosophy. On the other hand, it seems quite natural to say that, in the circumstances, what *U meant* (or part of what *U* meant) by making the statement he in fact made was that Mr. *X* is not much good (or that *U* thinks Mr. *X* is not much good) at philosophy. This is something *the utterer* implied by making the statement he did in this context, not something implied by the sentence uttered or by the statement *U* made by uttering the sentence.

The theory of conversation

With respect to what *U means* by a linguistic utterance, Grice proposed to separate what *U says* and what *U implicates* (e.g. implies, indicates, or suggests). What *U* says is to be closely tied to the conventional meaning of the words uttered, which both falls short and goes beyond what is said.

It falls short because a specification of what *U* said on a particular occasion must take into account not only the conventional meaning of the sentence used but also (e.g.) the references of referring expressions (e.g. proper names, demonstratives, and indexicals) and the time and place of utterance. *What U said* is to do duty for *what U stated* or *the proposition expressed* by *U*. Where the sentence uttered is of the type conventionally associated with the speech act of asserting (i.e. when it is in the "indicative mood") what is said will be straightforwardly *truth-conditional*. When the sentence uttered is in the imperative or interrogative mood, what is said will not be straightforwardly truth-conditional, but it will be systematically related to the truth conditions of what *U* would have said, in the same context, by uttering the indicative counterpart (or one of the indicative counterparts) of the original sentence.

The conventional meaning of a sentence also goes beyond what is said because of devices that signal the performance of "noncentral speech acts" parasitic upon the performance of the "central speech acts" of asserting, questioning, and ordering. Such devices, although they play a part in determining what *U meant*, play no part in determining what *U said*. If *U* utters (1) rather than (2),

1 She is poor but she is honest
2 She is poor and she is honest

very likely *U* will be taken to be implying that there is (or that someone might think there is) some sort of contrast between poverty and honesty (or her honesty and her poverty). This type of implication is no part of what *U* says because it does not contribute in any way to the *truth conditions* of the utterance. By uttering (1), *U* has *said* only that she is poor and she is honest; and this does not entail that there is any (e.g.) contrast between poverty and honesty (or between her poverty and her honesty). The implication in question Grice calls a *conventional implicature*.

According to Grice, by uttering (1) *U* is performing two speech acts, *saying* that she is poor and she is honest and *indicating* (or *suggesting*) that someone (perhaps *U*) has a certain attitude toward what *is* said. Grice did not develop this idea; he just left us with the claim that a conventional implicature is determined (at least in part) by the (con-

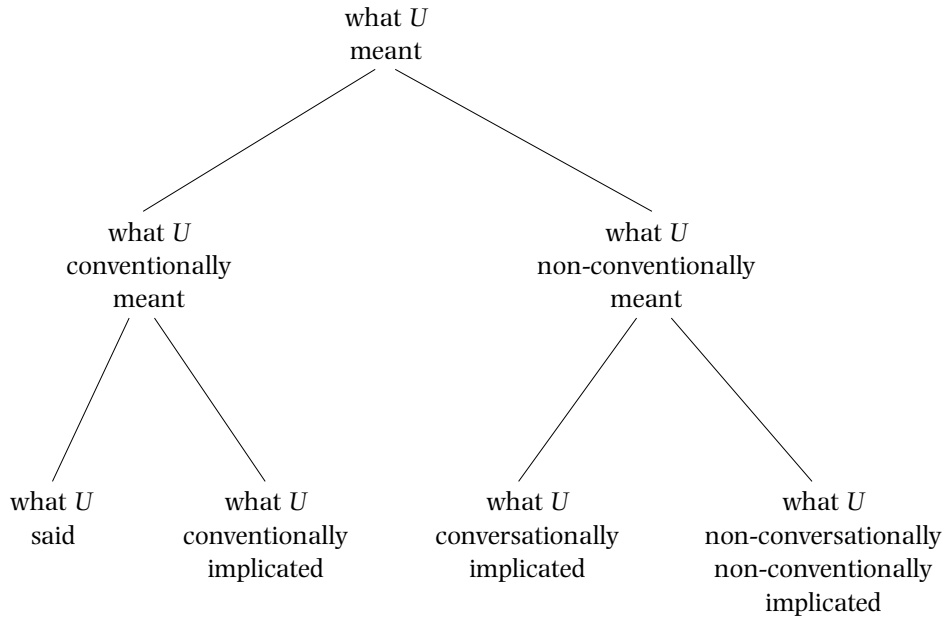


Figure 1

ventions governing) the words used. He does stress, however, that the sort of implication we have just been considering is not a *presupposition* (as originally defined by Strawson and adopted by others). B is a presupposition of A, just in case the truth or falsity of A requires the truth of B. (If the *truth* of A requires the truth of B, but the falsity of A does not, B is an *entailment* of A.) More precisely, if A presupposes B, A lacks a truth value if B is false. But as Grice points out, an utterance of (1) can be false even if the implied proposition is false, effectively scotching the idea that the implication is presupposition (at least not on the standard semantic conception of that notion). It is Grice's view that any alleged presupposition is either an entailment or an implicature.

For something to be (part of) what *U* says, it must also be (part of) what *U* meant, that is, it must be backed by a complex intention of the sort that forms the backbone of Grice's theory of meaning (see figure 1). If *U* utters the sentence "Bill is honest" ironically, on Grice's account *U* will *not* have said that Bill is honest: *U* will have *made as if to say* that Bill is honest. For it is Grice's view that a statement of the form "by uttering *x*, *U* said that *p*" entails the corresponding statement of the form "by uttering *x*, *U* meant that *p*." So on Grice's account, *one cannot unintentionally say something* (a fact that has interesting consequences for, for example, slips of the tongue and misused expressions).

Grice's work provides a breakdown of what *U* meant as shown in figure 1.

What *U* conventionally implicates and what *U* says are both closely tied to the conventional meaning of the sentence uttered, and they are taken by Grice as exhausting *what U conventionally means* (i.e. means by virtue of linguistic convention). Let us now

turn to what *U* non-conventionally means. Consider again, the example concerning Professor *U*'s evaluation of Mr. *X*. By uttering the sentence "Mr. *X* has excellent handwriting and is always very punctual," *U* said (or made as if to say) that Mr. *X* has excellent handwriting and is always very punctual. In addition, on Grice's account *U conversationally implicated* that Mr. *X* is not much good at philosophy (there is a *conversational implicature* to the effect that Mr. *X* is not much good at philosophy). Conversational implicature is a species of pragmatic (non-semantic, non-conventional) implication and is to be contrasted with the (at least partly semantic) implication that Grice calls conventional implicature. The principal difference between a conventional and a conversational implicature is that the existence of a conventional implicature depends upon the presence of some particular conventional device (such as "but," "moreover," "still," "yet," or heavy stress) whereas the existence of a conversational implicature does not.

Grice proposes to explain the possibility of a divergence between what *U* says and what *U* means (or at least between what *U* conventionally means and what *U* means) by appeal to the nature and purpose of rational interaction. Conversation is viewed by him as a characteristically purposeful and cooperative enterprise governed by what he calls:

The *Cooperative Principle*: Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. Subsumed under this principle, Grice distinguishes four categories of more specific maxims and submaxims enjoining truthfulness, informativeness, relevance, and clarity:

Quantity: Make your contribution as informative as is required (for the current purposes of the exchange); do not make your contribution more informative than is required.

Quality: Try to make your contribution one that is true. Specifically: (1) Do not say what you believe to be false; (2) Do not say that for which you lack adequate evidence.

Relation: Be relevant.

Manner: Be perspicuous. Specifically: (1) Be brief; (2) Be orderly; (3) Avoid ambiguity; (4) Avoid obscurity of expression.

Grice's basic idea is that there is a systematic correspondence between what *U* means and the assumptions required in order to preserve the supposition that *U* is observing the Cooperative Principle and conversational maxims. In the case of Professor *U*'s evaluation of Mr. *X*, on the surface the Cooperative Principle or one or more of the maxims is intentionally and overtly not fulfilled. By saying "Mr. *X* has excellent handwriting and is always very punctual" (in this particular context), *U* seems not to have fulfilled one of the maxims of Quantity or the maxim of Relation. (If Mr. *X* is one of *U*'s students, *U* must be in a position to volunteer more relevant information than judgments about Mr. *X*'s handwriting and timekeeping; furthermore, *U* knows that more information, or more relevant information, is required.) The hearer is naturally led to the conclusion that *U* is trying to convey something else, something more relevant to the purposes at hand. In the circumstances, if *U* thought Mr. *X* was any good at philosophy he would have said so. So *U* must think Mr. *X* is no good at philosophy and be unwilling to say so. And so *U* has conversationally implicated that Mr. *X* is no good at philosophy.

One interesting feature of this example is that it might well be the case that only what is implicated is meant (i.e. backed by *U*'s communicative intentions). *U* may have no idea what Mr. *X*'s handwriting is like because Mr. *X* has shown *U* only typed manuscripts of his work (or because he has never shown *U* anything), and *U* may have no opinion as to whether or not Mr. *X* is punctual. In such a version of the envisioned scenario, *U* has only *made as if to say* that Mr. *X* has excellent handwriting and is always very punctual because *U* had no intention of inducing (or activating) in his audience the belief that (*U* thinks that) Mr. *X* has excellent handwriting and is always very punctual. The truth-values of what *U* said (or made as if to say) and what *U* conversationally implicated may of course differ. Mr. *X* may have quite atrocious handwriting, and *U* may know this; but given the relevance of what is conversationally implicated, *U* may care very little about the truth-value of what he has said (or made as if to say). The primary message is to be found at the level of what is conversationally implicated.

In general, Grice claims, a speaker conversationally implicates that which he must be assumed to think in order to maintain the assumption that he is observing the Cooperative Principle (and perhaps some conversational maxims), if not at the level of what is said, at least at the level of what is implicated. At some overarching level of what is *meant*, *U* is presumed to be observing the Cooperative Principle. The wording of Grice's maxims suggests that some concern only what is said (e.g. "Do not say what you believe to be false") while others concern, perhaps, what is *meant* (e.g. "Be relevant"). We should probably treat this as something of an uncharacteristic looseness of expression on Grice's part. Except for the maxims under Manner (which can apply only to what is said) it seems reasonable to understand Grice as allowing a maxim not to be fulfilled at the level of what is said to be licensed or overridden by adherence at the level of what is implicated. On such a view, blatantly violating a maxim at the level of what is said but adhering to it at the level of what is implicated would not necessarily involve a violation of the Cooperative Principle.

Some important questions are still unanswered: How are *saying* and *implicating* to be defined? How are implicatures calculated? What is the status of the Cooperative Principle and maxims? What happens when a speaker cannot simultaneously observe all of the maxims? It is important to see how Grice attempts to face such questions.

No one should deny that in the example of the evaluation of Mr. *X* there is an intuitive and obvious distinction to be made between what *U* said and what *U* conversationally implicated. But in view of the sorts of example that really bother Grice – "the *F* is *G*," "*p* or *q*," "if *p* then *q*," etc. – he could not rest with an intuitive distinction. The example concerning the evaluation of Mr. *X* is clear-cut, obvious, and uncontentious. And herein lies the problem. The examples of purported conversational implicature that most interest Grice are philosophically important ones with respect to which many philosophers have not felt the need to invoke such a distinction. This might be because it is not at all obvious that there is such a distinction to be made in the cases in question (or if there is, how relevant it is), or because adherence to some form of the "meaning is use" dogma has blinded certain philosophers to the possibility of such a distinction. So Grice ultimately needs *analyses* of "what is said" and "what is conversationally implicated" in order to get philosophical work out of these notions.

Grice hopes to analyze the notion of saying in terms of utterers' intentions. This proposal will be examined after discussion of Grice's theory of meaning. He attempts to define conversational implicature in terms of an, as yet, undefined notion of saying. The following schema is supposed to be a first step:

Someone who, by (in, when) saying (or making as if to say) that *p* has implicated that *q*, has conversationally implicated that *q*, provided that (1) he is to be presumed to be observing the conversational maxims, or at least the Cooperative Principle; (2) the supposition that he is aware that *q* is required in order to make his saying or making as if to say *p* consistent with this presumption; and (3) the speaker thinks (and would expect the hearer to think that the speaker thinks) that it is within the competence of the hearer to work out, or grasp intuitively, that the supposition mentioned in (2) is required.

We appear to have here a set of *necessary* conditions. The conditions are not *sufficient* because conventional implicatures are not excluded. Whenever there is a conversational implicature, one should be able to reason somewhat as follows: (i) *U* has said that *p*; (ii) there is no reason to suppose that *U* is not observing the Cooperative Principle and maxims; (iii) *U* could not be doing this unless he thought that *q*; (iv) *U* knows (and knows that I know that *U* knows) that I can see that *U* thinks the supposition that *U* thinks that *q* is required; (v) *U* has done nothing to stop me thinking that *q*; (vi) *U* intends me to think, or is at least willing to allow me to think, that *q*; (vii) and so, *U* has implicated that *q*. In each of the cases Grice considers, it does seem to be possible to justify the existence of the implicature in question in this sort of way. But notice that *q* is simply introduced without explanation in step (iii), so Grice has certainly not stated any sort of method or procedure for calculating the content of conversational implicatures. A good deal of work needs to be done on the calculation of particular implicatures if Grice's evident insights are to form the basis of a finally acceptable theory.

A necessary condition on conversational implicatures that is intimately connected to condition (3) is that they are *intended*. This follows, if not from condition (3), at least from the fact that (a) what *U* implicates is part of what *U* means, and (b) what *U* means is determined by *U*'s communicative intentions. A hearer may think that, by saying that *p*, *U* has conversationally implicated that *q* (*A* may even have reasoned explicitly in the manner of (i)–(vii) above). But if *U* did not intend the implication in question it will not count as a conversational implicature.

We have, then, four conditions that are necessary but not sufficient for classifying an implication as a conversational implicature. Entailments do not seem to have been excluded. In order that we may stay focused on the relation between the speaker and certain propositions, let us make a harmless addition to Grice's terminology. If the proposition that *p* entails the proposition that *q*, then if *U* is a competent speaker who says that *p, U* thereby says* that *q*. So if *U* is a perfectly competent English speaker who has sincerely uttered the sentence "John is a bachelor," not only has *U* said (and said*) that John is a bachelor, he has also said* that John is unmarried. It seems desirable that no proposition be both an entailment and a conversational implicature of the same utterance. But it is not obvious that the conditions laid down thus far on conversational implicature actually rule out entailments. Furthermore, Grice cannot just impose a further condition to the definition to the effect that no entailment is a conversational

implicature. One of Grice's avowed aims is to ward off certain ordinary language arguments by invoking a sharp distinction between what we are now calling conversational implicature and entailment; so it is not good enough for him to use the notion of an entailment in a definition of conversational implicature.

A fifth condition Grice imposes on conversational implicatures seems to help. Unlike an entailment, a conversational implicature is supposed to be *cancelable* either explicitly or contextually, without contradiction. If *U* says* that *p*, and *p* entails *q*, then *U* cannot go on to say* that not-*q* without contradiction. For example, *U* cannot say "John is a bachelor and John is married." But if *U* says* that *p*, and thereby conversationally implicates that *q*, *U* can go on to say* that not-*q* without contradiction. Consider again the case of *U*'s evaluation of Mr. *X*. After uttering "Mr. *X* has excellent handwriting and is always very punctual," *U* might (without irony) continue "Moreover, Mr. *X*'s recent modal proof of the immortality of the soul is a brilliant and original contribution to philosophy." In the light of the first comment, this addition might be rather odd, but it would not result in *U* contradicting himself. (In addition to distinguishing conversational implicatures from entailments, the cancelability test is also supposed to distinguish conversational from conventional implicatures. Although it will not lead to contradiction, attempting to cancel a conventional implicature will result in a genuinely linguistic transgression of some sort. This is precisely because there is a distinct semantic component to conventional implicatures.)

Putting these five conditions together, we come as close as we can with Grice's machinery to a set of necessary and sufficient conditions on conversational implicature.

Philosophical psychology

For Grice, the principles involved in an account of conversational implicature are to be grounded in a philosophical psychology that explicates the purportedly hierarchical relationships that hold between the various types of psychological states we ascribe to creatures that can reason and form complex intentions. The beginnings of this line of thought can be traced to the end of his 1957 paper "Meaning." It contains the seeds of (1) the view that the Cooperative Principle and conversational maxims (in particular the maxim enjoining relevance) are to play a central role not only in an account of possible divergences between what *U* said and what *U* meant but also in an account of the resolution of ambiguities, and (2) the view that the use of language is one form of rational activity and that the principles at work in the interpretation of linguistic behavior are (or are intimately related to) those at work in interpreting intentional *non*-linguistic behavior.

Two questions spring to mind immediately: (1) What are the relative rankings of the maxims in cases where it is hard (or impossible) for *U* to observe all of them (or all of them to the same degree), and why? (2) What is the basis for the assumption that speakers will in general (*ceteris paribus* and in the absence of indications to the contrary) proceed in the manner prescribed by the Cooperative Principle and maxims?

Grice is explicit about the position of at least one of the maxims of Quality in any hierarchy. Suppose *A* is planning an itinerary for a vacation to France. *A* wants to see his friend *C*, if so doing would not require too much additional traveling. *A* asks *B*

"Where does *C* live?" *B* replies "Somewhere in the south of France." *B* knows that *A* would like more specific information but he is not in a position to be more specific. So *B* is faced with not fulfilling either a maxim of Quality or a maxim of Quantity. Quality wins out. The maxims of Quality have a very special status within Grice's overall theory and Grice entertains the idea that the first maxim of Quality should be part of some broader background; the other maxims come into operation only on the assumption that the maxim of Quality is satisfied. The maxims of Quality (or at least the first maxim of Quality) should not be thought of as admitting of degree or varying across cultures. In some sense this is an empirical matter; but unlike the maxims of Quantity and Manner, it does not seem very plausible to suppose that there are thriving cultures in which standardly people do not behave (for particular reasons to be determined by anthropologists) as if they are observing the maxims of Quality.

Grice was not satisfied with the idea that it is just a well-recognized empirical fact that people do behave in accordance with the maxims and the Cooperative Principle, that in childhood they learned to do so and have not lost the habit. He wanted to find a basis that underlies our behavior and believed it would have a moral dimension: not only do we *in fact* behave in the required way, but it is *reasonable* for us to do so, and the practice is something we *should not* abandon given our common purposes or goals. Conversation is one among a range of forms of rational activity for Grice. Observance of the CP and maxims is reasonable (rational): anyone concerned about the goals central to communication must be expected to have an interest, given suitable circumstances, in participation in informational exchanges that will be profitable only on the assumption that they are conducted in general accordance with the CP and maxims.

On Grice's view, value predicates such as "proper," "correct," "optimal," and "relevant" cannot be kept out of an account of rational activity because a rational creature is essentially a creature that *evaluates*. Whether a value-oriented approach to the interpretation of intentional behavior can be developed in a fruitful way remains to be seen. But as Grice's unpublished work on ethics and philosophical psychology becomes more widely available, there will likely be a resurgence of interest in the matter of the precise location of the theory of conversation within a larger scheme.

The logic of natural language

One task of semantics is to provide a systematic characterization of judgments concerning truth, falsity, entailment, contradiction, and so on. In the light of theoretical considerations, an initial judgment of, say, entailment might be rejected on the grounds that the perceived implication is an implicature rather than an entailment. So far, we have considered only examples of what Grice calls "particularized" conversational implicature, examples in which there is no temptation to say that the relevant implication is an entailment (or a "presupposition"). Of more philosophical interest are "generalized" conversational implicatures, the presence and general form of which depend little upon the particular contextual details. Examples discussed by Grice include those attaching to utterances of sentences containing intentional expressions like "look," "feel," and "try," and "logical" expressions such as "and", "or," "if," "every," "a," and "the."

According to Grice, philosophers who see divergences in meaning between "formal devices" such as "&," "$\vee$," "$\supset$," "$(\forall x)$," "$(\exists x)$," and "(ιx)" and their natural language counterparts tend to belong to one of two camps, which he calls "formalist" and "informalist." The informalist position is essentially the one taken by Strawson (and others of the "ordinary-language movement"). The formalist camp is dominated by positivists and others who view natural language as inadequate to the needs of the science and philosophy of an age of precision. A typical formalist recommends the construction of an "ideal" or "logically perfect" language such as the language of first-order quantification theory with identity (or some suitable extension thereof). Since the meanings of the logical particles are perfectly clear, using an ideal language, philosophers can state propositions clearly, clarify the contents of philosophical claims, draw the limits of intelligible philosophical discourse, draw the deductive consequences of sets of statements, and generally determine how well various propositions sit with each other.

Grice views the formalists and informalists as mistaken in the assumption of semantic divergence. Both sides have taken mere pragmatic implications to be parts of the meanings of sentences of natural language containing "logical" expressions. The case of "and" highlights some important methodological considerations. Although it is plausible to suppose that "and" (when it is used to conjoin sentences) functions semantically just like "&," there are certainly sentences in which it appears to function rather differently:

1 Jack and Jill got married and Jill gave birth to twins.
2 Nero yelled and the prisoner began to tremble.

Someone who uttered (1) would typically be taken to imply that Jack and Jill got married *before* Jill gave birth to twins. And someone who uttered (2) would typically be taken to imply that Nero's yelling contributed in some way to the prisoner's trembling. Thus one might be led to the view that "and" is not always understood as "&," that it is (at least) three ways ambiguous between truth-functional, temporal, and causal readings.

The postulation of semantically distinct readings looks extravagant and Grice suggests it is good methodological practice to subscribe to "modified Occam's razor": *senses are not to be multiplied beyond necessity*. Given the viability of the distinction between what is said and what is meant, if a pragmatic explanation is available of why a particular expression appears to diverge in meaning in different linguistic environments (or in different conversational settings) then *ceteris paribus* the pragmatic explanation is preferable to the postulation of a semantic ambiguity. Grice's idea is that the implication of temporal sequence attaching to an utterance of (1) can be explained in terms of the fact that each of the conjuncts describes an event (rather than a state) and the presumption that *U* is observing the submaxim of Manner enjoining orderly deliveries. It seems to be Grice's view, then, that by uttering (1) *U* will conversationally implicate (rather than say) that Jack and Jill got married before Jill gave birth to twins (if this is correct then what is conversationally implicated would appear to entail what is said in this case). Similarly, the implication of causal connection attaching to an utterance of (2) is apparently to be explained in terms of the presumption that the speaker is being relevant. Before looking at problems for this proposal, I want first to get clear about its strengths.

Conversational explanations trump semantic ambiguities on grounds of theoretical economy and generality. A conversational explanation is free: the mechanisms appealed to are already in place and independently motivated. The generality lost by positing several readings of "and" is quite considerable. First, implications of (e.g.) temporal priority and causal connection attach to uses of the counterparts of "and" across unrelated languages. Second, implications of the same sorts would surely arise even for speakers of a language containing an explicitly truth-functional connective "&." Third, the same implications that attach to utterances of "*p* and *q*" would attach to an utterance of the two sentence sequence "*p. q*" not containing an explicit device of conjunction. On *methodological* grounds, then, pragmatic accounts of the temporal and causal implications in (1) and (2) are preferable to accounts that appeal to semantic ambiguity.

Grice opposes postulating idiosyncratic *pragmatic rules* with which to derive generalized implicatures. Conversational implicatures must be explicable in terms of the Cooperative Principle and maxims, construed as *general* antecedent assumptions about the rational nature of conversation. To call an implicature "generalized" rather than "particularized" is only to acknowledge the fact that the presence of the implicature is relatively independent of the details of the particular conversational context, a fact that is to be explained by the cooperative nature of conversation.

A second challenge to classical logic semantics came from Strawson, who challenged Russell on the grounds that the theory does not do justice to ordinary usage: speakers use descriptions to *refer*, not to quantify, and hence Russell's theory is open to a number of objections (see STRAWSON). But according to Grice, a number of Strawson's objections can be defused by distinguishing sentence meaning, what is said, and what is meant.

In Grice's terminology, one of Strawson's main complaints against Russell is that his theory conflates the meaning of a sentence "the *F* is *G*" and what *U* says by uttering this sentence (and similarly the subsentential counterparts of these notions) and so cannot explain the fact that *U* may say different things on different occasions by uttering the same sentence. Grice is right that Strawson can get no mileage out of Russell's failure to separate sentence meaning and what is said in his discussions. Upon reflection it is clear that Russell's concern is with what is said rather than sentence meaning. If Russell were being more precise, he would not say that the *sentence* "the *F* is *G*" is equivalent to the *sentence* "there is exactly one *F* and every *F* is *G*"; rather, he would say that what *U* says by uttering "the *F* is *G*" on a particular occasion is that there is exactly one *F* and every *F* is *G* (occurrences of "*F*" in the foregoing may, of course, be elliptical). The fact that a description (or any other quantified noun phrase) may contain an indexical component ("the *present* king of France," "every man *here,*" etc.) does not present a problem: all this means is that there are some descriptions that are subject to the Theory of Descriptions (see RUSSELL) *and* a theory of indexicality. Grice is surely right, then, that although we need a sharp distinction between sentence meaning and what is said (and their subsentential counterparts), Strawson's appeal to this distinction when challenging Russell is empty.

Grice neatly disposes of the view that descriptions are ambiguous between Russellian and referential (or identificatory) readings. When a description is used to identify something, what *U* means diverges from what *U* says. What *U* says is given by the Russellian

expansion but *U* also intends to communicate information about some particular individual, and although this is part of what *U* means, it is not part of what *U* says. This provides a perfectly satisfactory account of what is going on when *U* uses a description that does not fit its target, but such cases are not needed to see Grice's distinctions at work. According to Grice, when a description is used in an identificatory way, there will *always* be a mismatch between what *U* says and what *U* means (even where the description uniquely fits the individual the speaker intends to communicate information about) because what is said is, on Russell's account, analyzable as a quantificational proposition, whereas what is meant will always include a singular or object-dependent proposition.

Again, methodological considerations strongly favor the Gricean account of referential usage over an account that posits a semantic ambiguity: (1) If we were taught explicitly Russellian truth conditions, referential usage would still occur; (2) exactly parallel phenomena occur with indefinite descriptions and other quantified noun phrases; (3) modified Occam's razor enjoins us to opt for the simpler of two theories, other things being equal. Subsequently, far more detailed defenses of Russell along Gricean lines have been proposed by other philosophers, but the debts these works owe to Grice are considerable. More generally, a debt is owed to Grice for rejuvenating the position that classical logic is a remarkably useful tool as far as the semantics of natural language is concerned.

The theory of meaning

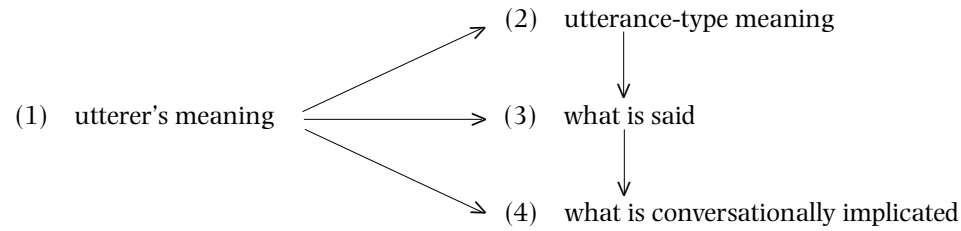
Grice attempted to analyse or explicate *what is said* and *what is implicated* in terms of intention, belief, desire, and recognition. Analyzing locutions of the forms "*X* did *Y* intentionally," "*X* caused *Y*," "*X* is true," "*X* entails *Y*," and so on, has been seen by many philosophers as a central task of philosophy. Grice's analyses of "by uttering *X*, *U* meant that *p*," "*X* means '*p*'," and "by uttering *X*, *U* said that *p*" seem to have a reductive and explicative flavor in that it appears to be his view that locutions of the forms can be wholly explicated without appealing to semantical concepts.

He begins with what *people* mean rather than with what this or that expression, sign, or action means, seeking to analyze this in terms of complex audience-directed intentions on the part of the utterer, and to analyze utterance-type meaning (e.g. sentence meaning and word meaning) in terms of utterer's meaning.

Although Grice aims to neutralize many ordinary language maneuvers with his saying/implicating distinction, one of the driving forces behind his work is still the idea that the meaning of an expression is a function of what its users do with it. Abstracting away from certain details that I will get to later, the direction of analysis for Grice is shown in figure 2.

The idea, then, is to begin by providing an analysis of (1) utterer's meaning, and then to use this analysis in an analysis of (2) utterance-type meaning. (3) What is said is then to be defined in terms of a near coincidence of utterer's meaning and utterance-type meaning (for certain utterance-types); and finally (4) conversational implicature is to be defined in terms of saying and utterer's meaning.

Although Grice does not address this point directly, it is clear that the task of explicating the locution "by uttering *x*, *U* said that *p*" takes on some urgency for him,



The "__ → __" is understood as "__ (or its analysis) plays a role in the analysis of __ (but not vice versa)."

Figure 2

because the saying/implicating distinction is so central to his attempts to counter ordinary language arguments of the sort examined earlier. A direct *analysis* of saying appears out of the question because Grice openly declares that he is using "say" in a special sense, and this precludes systematic appeal to intuitions about ordinary usage. By contrast, when it comes to pronouncing on the truth of instances of "by uttering *x*, *U meant* that *p*," Grice believes he can help himself to such intuitions, many of them quite subtle. Strictly speaking, then, saying is to be *defined* rather than analyzed.

To some philosophers and linguists, Grice's program seems to constitute something of a snub to serious compositional semantics. The idea that sentence meaning is to be analyzed in terms of utterer's meaning has been felt to conflict with (1) the fact that knowing the meaning of a sentence is typically a necessary step in working out what *U* meant by uttering that sentence, i.e. for recovering *U*'s communicative intentions, and (2) the fact that the meaning of a sentence is determined, at least in part, by the meanings of its parts (i.e. words and phrases) and the way the parts are put together (syntax). Both of these charges are based on misunderstandings of Grice's project, as will become clear.

Utterer's meaning

The basic Gricean analysis of utterer's meaning is this:

I. "By uttering *x*, *U* meant something" is true iff for some audience *A*, *U* uttered *x* intending:
   (1) *A* to produce some particular response *r*,
   (2) *A* to recognize that *U* intends (1), and
   (3) *A*'s recognition that *U* intends (1) to function, in part, as a reason for (1).
   To provide a specification of *r*, says Grice, is to say *what U* meant. Where *x* is an "indicative" utterance, *r* is *A's believing something*.

II. "By uttering *x*, *U* meant that *p*" is true iff for some audience *A*, *U* uttered *x* intending:
   (1) *A* to believe that *p*,
   (2) and (3) as above.

This type of complex intention Grice calls an "*M*-intention": by uttering *x*, *U* meant that *p* iff for some audience *A*, *U* uttered *x* *M*-intending *A* to believe that *p*.

Two general problems face II. The first is that Grice provides a number of examples in which it would be correct to say that *U* means that *p* but incorrect to say that *U* intends *A* to believe that *p* (1989: 105–9). Suppose *U* is answering an examination question and says "The Battle of Waterloo was fought in 1815." Here *U* meant that the Battle of Waterloo was fought in 1815; but *U* did not *M*-intend the examiner to think that the Battle of Waterloo was fought in 1815 (typically, *U* will be under the impression that the examiner already knows the answer). In response to this and related examples, Grice suggests that clause (1) of II. be changed to (1_1):

(1_1) *A* to think that *U* thinks that *p*.

A distinction is then made between *exhibitive* utterances (utterances by which *U M*-intends to impart the belief that *U* has a certain propositional attitude) and *protreptic* utterances (utterances by which *U M*-intends, *via* imparting a belief that [*U*] has a certain propositional attitude, to induce a corresponding attitude in the hearer).

The suggested revision may not seem to comport with the commonly held view that the primary purpose of communication is the transfer of information about the world; on the revised account, the primary purpose seems to be the transfer of information about one's mental states. Another worry is that even if the proposed revision is an improvement, it does not weaken the analysis in such a way as to let in cases of reminding (some cases of which bring up another problem). Suppose *U* knows that *A* thinks that *p* but needs reminding. So *U* does something by which he means that *p*. Not only does it seem incorrect to say (as the original analysis would require) that *U* intends *A* to think that *p* – *U* knows that *A* already thinks that *p* – it also seems incorrect to say (as the modified analysis requires) that *U* intends *A* to think that *U* thinks that *p* (*U* may know that *A* already thinks that *U* thinks that *p*). What seems to be needed here, says Grice, is some notion of an *activated* belief: (1) needs to be changed not to (1_1) but to something more like (1_2):

(1_2) *A* actively to believe that *U* thinks that *p*.

But there seems still to be a problem involving reminding. Suppose *A* has invited *B* over for dinner tonight at seven-thirty. *B* has agreed to come but *U* doubts *B* will show up and says as much to *A*. At seven o'clock, *U* and *A* are deep in philosophical conversation and *U*, realizing that *A* has lost track of time, says "*B* will be here in half an hour." This type of example suggests we are better off with something like (1_3), at least for some cases:

(1_3) *A* actively to believe that *p*.

So perhaps a disjunctive clause is going to be required in any finally acceptable analysis.

Perhaps the problem with the first clause of II. is an instance of a more general difficulty concerning the content of the intention (or *M*-intention) characteristic of communicative behavior. This seems to be the view of Searle (1969). One way of putting Searle's general point is as follows: by paying too much attention to examples in which *U* intends to induce in *A* some propositional attitude or other, Grice has mistakenly taken a particular type of intention that does in fact accompany many utterances – the

subintention specified in clause (1) – to be an essential ingredient of communicative behavior. But there are just too many cases of meaning involving linguistic (or otherwise conventional) utterances in which *U* does not seek to induce in an audience any propositional (or affective) attitude. Searle brings up three problems: first, it is not at all clear what attitude I *M*-intend to impart when making a promise by uttering a sentence of the form "I promise to __"; second, sometimes I don't care whether I am believed or not; I just feel it is my duty to speak up; third, only an egocentric author intends me to believe that *p* because he has said so.

These are genuine difficulties for Grice's analysis as it stands, but they do not seem to warrant abandoning Grice's project; rather they suggest that the specification of the *type* of response mentioned in the first clause needs to be weakened to something like the following:

(1_4) *A* actively to entertain the belief/thought/proposition that *p*.

Of course, in many cases *U* also intends (or at least would like) *A* to go on to believe that *p*, but this fact would not enter into the analysis of utterer's meaning. A revision along these lines might provide the beginning of a way out of Searle's problems.

The second problem is that clause (3) of II. seems problematic. The original motivation for clause (2) is clear. It is not enough, Grice points out, for *U* to mean that *p*, that *U* utter *x* intending *A* to think that *p*. *U* might leave *B*'s handkerchief near the scene of the murder with the intention of getting the detective (actively) to entertain the thought that *B* is the murderer. But there is no temptation to say that by leaving the handkerchief, *U meant* that *B* is the murderer. Hence clause (2), which requires *U* to intend *A* to recognize the intention specified in the first clause (however stated).

But what of clause (3)? Grice wants this in order to filter out cases in which some natural feature of the utterance makes it *completely obvious* that *p*. He worried about cases like this: in response to an invitation to play squash, Bill displays his bandaged leg. According to Grice, we do not want to say that Bill *meant* that his leg was bandaged (though we might want to say that he meant that he could not play squash, or even that he had a bad leg).

Many people's intuitions are less robust than Grice's here. He seems to be worried that in cases like these there is something approximating natural meaning that interferes with the idea of Bill non-naturally meaning that he has a bandaged leg. Given the links Grice seeks to forge between natural and non-natural meaning, it is not clear why the putative presence of natural meaning is supposed to be problematic, and so it is not clear why the third clause of II. is needed. Grice himself brings up cases that seem to create a problem for the third clause. Suppose the answer to a certain question is "on the tip of *A*'s tongue." *U* knows this; that is, *U* knows that *A* thinks that *p* but can't quite remember. So *U* reminds *A* that *p* by doing something by which he (*U*) means that *p*. In such a scenario, even if *U* has the intention specified in the first clause (however stated), it does not seem to be the case that *U* has the intention specified in the third clause. It is noteworthy that the examples Grice uses to justify the third clause involve non-linguistic utterances (Grice's "John the Baptist" and "bandaged leg" cases). However, it is possible to construct cases involving properly linguistic utterances in which the fact that *p* is made just as obvious by the utterance as in Grice's non-linguistic cases. Consider an utterance by me of (e.g.) "I'm right here" yelled in the direction of

someone known to be looking for me. Here there is a strong inclination to say that I did not *mean* what I said.

Problems await Grice if he does not concede the third clause is overly restrictive. Ultimately, he wants to define locutions of the form "by uttering *x*, *U* said that *p*"; but one of the conjuncts in his proposed definiens is "by uttering *x*, *U* meant that *p*." So if he refuses to allow that (e.g.) I can mean that I can speak in a squeaky voice by uttering, in a squeaky voice, "I can speak in a squeaky voice," Grice will be forced either to conclude that I have not *said* that I can speak in a squeaky voice, or else to abandon the idea of defining saying in terms of utterer's meaning (he cannot, of course, say that in such a scenario I have only "made as if to say" that I can speak in a squeaky voice). It would seem, then, that the third clause will have to be discarded (or at least modified) if saying requires meaning.

One positive result of discarding the third clause would be the disappearance of the "tip-of-the-tongue" problem. Another would be that Bill could mean that he had a bandaged leg in the scenario above, which is not obviously incorrect. When it comes to linguistic utterances, there might well be another interesting consequence. Typically, linguistic utterances do not seem to be underwritten by intentions as complex as *M*-intentions. Weakening the analysans by the removing clause (3) goes a long way toward quieting this worry; however, there are grounds for thinking that the relevant intention will have to be more complex than the one specified by clauses (1) and (2).

The following type of example shows that clauses (1), (2), and (3) do not specify a rich enough intention (or batch of intentions). Suppose *A*, a friend of mine, is about to buy a house. I think the house is rat-infested, but I don't want to mention this outright to *A* so I let rats loose in the house knowing that *A* is watching me. I know that *A* does not know that I know that he is watching me do this. I know *A* will not take the presence of my rats to be natural evidence that the house is rat-infested; but I do know, indeed I intend, that *A* will take my letting rats loose in the house as grounds for thinking that I intend to induce in him the belief that the house is rat-infested. Conditions (1)–(3) of II. above are fulfilled. But surely it is not correct to say that by letting rats loose in the house I mean that the house is rat-infested.

The problem is that in this example my intentions are not *wholly overt*. One possible remedy involves adding a fourth clause:

(4) *A* to recognize that *U* intends (2).

But the same sort of counterexample can still be generated, and then we need a fifth clause, then a sixth, and so on. Grice proposed to block an infinite regress by adding a condition that would prohibit any "sneaky" intention: instead of adding additional clauses, his idea was to add a second part to the analysis, the rough import of which is that *U* does not intend *A* to be deceived about *U*'s intentions (1)–(3). As long as *U* does not have a deceptive intention of this sort, *U* is deemed to mean that *p*.

Something like the following is best seen as the characterization of utterer's meaning that Grice left us to explore and refine:

III. By uttering *x*, *U* meant that *p* iff for some audience *A*,

(1) *U* uttered *x* intending *A* actively to entertain the thought that *p* (or the thought that *U* believes that *p*)
(2) *U* uttered *x* intending *A* to recognize that *U* intends *A* actively to entertain the thought that *p*
(3) *U* does not intend *A* to be deceived about *U*'s intentions (1) and (2).

Sentence meaning and saying

The idea of using utterer's meaning to explicate sentence meaning is thought by some philosophers to conflict with the idea that the meaning of a sentence is a function of the meanings of its parts (i.e. words and phrases) and their syntactical organization. Grice's project gets something "backwards" it is claimed: surely any attempt to model how we work out what someone means on a given occasion will progress from word meaning plus syntax to sentence meaning, and from sentence meaning plus context to what is said, and from what is said plus context to what is meant. And this clashes with Grice's view that sentence meaning is analyzable in terms of utterer's meaning.

But this is incorrect. Suppose there is a sentence *Y* of a language *L* such that *Y* means (pre-theoretically speaking) "Napoleon loves Josephine" (e.g. if *L* is English, then the sentence "Napoleon loves Josephine" will do). When *L*-speakers wish to mean that Napoleon loves Josephine they are more likely to use *Y* than a sentence Z that means (pre-theoretically speaking) "Wisdom is a virtue." To say this is not to say that it is *impossible* for *U* to mean that Napoleon loves Josephine by uttering Z, it's just to say that normally (usually, typically, standardly) *U* has a much better chance of getting across the intended message by uttering *Y*. Thus it might be suggested that an arbitrary sentence *X* means (in *L*) "Napoleon loves Josephine" iff (roughly) by uttering *X*, optimally, *L*-speakers mean (would/should mean) that Napoleon loves Josephine.

Grice is not committed to the absurd position that a hearer must work out what *U* meant by uttering a sentence *X* in order to work out the meaning of *X*. To see this as a consequence of Grice's theory is to ignore the connection between the theory of conversation and the theory of meaning. It is Grice's view that typically the hearer must establish what *U* has said (or made as if to say) in order to establish what *U* meant; and it is by taking into account the nature and purpose of rational discourse that the hearer is able to progress (via, for example, conversational implicature) from what *U* has said (or made as if to say) to what *U* meant. An *analysis* of sentence meaning in terms of utterer's intentions does not conflict with this idea.

We must distinguish (1) accounts of what *U* said and what *U* meant by uttering *X* and (2) accounts of how hearers recover what *U* said and what *U* meant by uttering *X*. What *U* meant by uttering *X* is determined solely by *U*'s communicative intentions; but of course the *formation* of genuine communicative intentions by *U* is constrained by *U*'s expectations: *U* cannot be said to utter *X* *M*-intending *A* to *ø* if *U* thinks that there is very little or no hope that *U*'s production of *X* will result in *A* *ø*-ing. If *U* *M*-intends *A* actively to entertain the belief that (*U* thinks) Napoleon loves Josephine, and *U* and *A* are both English speakers, *U* may well utter the English sentence "Napoleon loves Josephine." To say this is not to commit Grice to the view that sentences that are not directly (or *so* directly) connected to the proposition that Napoleon loves Josephine may not be employed to the same effect.

On the contrary, the theory of conversation is supposed to provide an explanation of how this is possible (in the right circumstances). On the assumption that *U* and *A* are both operating in accordance with the Cooperative Principle and maxims, there may well be facts about the context of utterance, the topic of conversation, background information, and so on that make it possible for *U* to mean that Napoleon loves Josephine by uttering a very different sentence. *U*'s conception of such things as the context of utterance, the topic of conversation, background information, and *A*'s ability to work out what *U* is up to may all play roles in the *formation* of *U*'s intentions; but this does not undermine the view that what determines what *U* means are *U*'s communicative intentions.

We can put aside, then, the question of the conceptual coherence of Grice's analytical program; the interesting questions concern the adequacy of his concrete proposals for explicating sentence meaning and saying. The basic idea is to analyze sentence meaning in terms of utterer's meaning, and then define saying in terms of a near coincidence of utterer's meaning and sentence meaning. Sentence meaning for Grice is a species of complete utterance-type meaning, the relevant analysandum for which is "*X* means '*p*'," where *X* is an utterance type and *p* is a specification of *X*'s meaning. Grice puts forward the following as indicative of the general approach he is inclined to explore:

IV. For population group *G*, complete utterance-type *X* means "*p*" iff (a) at least some (many) members of *G* have in their behavioral repertoires the procedure of uttering a token of *X* if they mean that *p*, and (b) the retention of this procedure is for them conditional on the assumption that at least some (other) members of *G* have, or have had, this procedure in their repertoires.

For a language containing no context-sensitive expressions, the technical difficulties involved in Grice's use of the variable "*p*" both in and out of quotes can be remedied easily enough. But once we turn (as we must) to complete utterance-type meaning for a language that contains indexicals such as "I" and "you," demonstratives such as "this" and "that," and anaphoric pronouns such as "him" and "her," it is clear that some work is needed to transform IV. into something acceptable. This is, I think, a very serious matter; for without such a transformation, Grice simply will not be able to provide an analysis of utterance-type meaning for a language like English, and consequently he will not be able to provide the sort of definition of saying he wants.

What, then, is the precise relation between sentence meaning and saying for Grice? It might be thought that if we abstract away from the problems raised by indexicals and other expressions that highlight the gap between sentence meaning and what is said, we will be able to move directly from *when uttered by U, X meant "p"* to *by uttering X, U said that p*. But there are two problems here. First, only where an utterance-type has certain features do we want to say that a token of that type may be used to *say* something. A motorist does not say anything, in Grice's sense, when he indicates an intention to make a left turn by signalling.

Second, certain cases involving, for example, irony or conversational implicature can be used to show that we cannot make the relevant move directly. If *U* utters the sentence "Smith is an honest man" ironically, although it would be true to say that the sentence in question means "Smith is an honest man," it would not be true to say that

U is *saying* that Smith is an honest man. On Grice's account, since *U* does not *mean* that Smith is an honest man (*U* has no intention of getting *A* to believe that (he believes that) Smith is an honest man) *U* is only *making as if to say* that Smith is an honest man. (Parallel remarks could apply in the case of Professor *U*'s utterance of the sentence "Mr *X* has wonderful handwriting and is always very punctual.") On Grice's account, what is said is to be found in the area where sentence meaning and utterer's meaning overlap. Abstracting away from context-sensitive expressions once again, it looks as though something like the following preliminary definition is on the right track:

V. By uttering *X*, *U* said that *p* iff
   (1) by uttering *X*, part of what *U* meant was that *p*
   (2) *X* consists of a sequence of elements (such as words) ordered in a way licensed by a system of rules (syntactical rules), and
   (3) *X* means "*p*" in virtue of the particular meanings of the elements in *X*, their order and their syntactical structure.

Grice's unhappiness with V. derives from the existence of *conventional* implicatures. Recall that Grice does not want to allow the sorts of implications that result from the use of words such as "but," "yet," "still," "even," and "moreover," to count as part of what is said. For example, if *U* (sincerely and non-ironically) utters the sentence "She is poor but she is honest," *U* does not *say* that there is some sort of contrast between poverty and honesty (or between her poverty and her honesty). Rather, *U* performs a "central speech act," by which *U* says that she is poor and she is honest, and performs in addition a "noncentral speech act," by which *U* conventionally implicates some sort of attitude toward what is said. Putting together what *U* says and what *U* conventionally implicates we get *what U conventionally means* (see figure 1). So for Grice, at best the three conditions in V. define *by uttering X, U conventionally meant that p* rather than *by uttering X, U said that p*.

This is as far as Grice goes. He leaves us with the non-trivial task of separating what *U* says and what *U* conventionally implicates, a rather disappointing terminus. The notion of what is said is for Grice a fundamentally important notion in philosophy. If this or that philosopher is unclear about what he is saying (as opposed to what he or she is implicating) then that philosopher is liable to make all sorts of mistakes, as is borne out, Grice thinks, by the crude way in which, for instance the causal theory of perception and the theory of descriptions have been written off by philosophers concerned with the nuances of ordinary language. Furthermore, not until what is said and what is conventionally implicated are separated can what is *conversationally implicated* be defined in the manner examined earlier.

So for Grice, an analysis of saying takes on some urgency, and it is unfortunate that he does not get any closer to one than he does in producing V. above. However, it may well be that Grice has brought us as far as we can go without crossing our own paths. Recall that he wants what is said to comprise the truth-conditional content of what is conventionally meant by someone making a statement; but he cannot appeal directly to truth conditions for fear of undermining one part of his project. There may be no simple way out of this. At the same time, only one part of Grice's project is threatened: the possibility of providing a definition of saying in terms of utterance-type meaning

and what is meant. No appeal to truth-conditional content is needed in analyses of *utterer's meaning* or *utterance-type meaning*, and to that extent Grice has certainly illuminated these important notions. In so doing, he has also alerted us to a host of important distinctions that philosophers, linguists, cognitive scientists, and literary theorists ignore at their peril.

Bibliography

Works by Grice

1989: *Studies in the Way of Words*, Cambridge: Cambridge University Press.
1991: *The Conception of Value*, Oxford: Clarendon Press.

Works by other authors

Avramides, A. (1989) *Meaning and Mind*, Cambridge, MA: MIT Press.
Grandy, R. E. and Warner, R. (eds.) (1986) *Philosophical Grounds of Rationality: Intentions, Categories, Ends*, Oxford: Clarendon Press.
Martinich, A. P. (1984) *Communication and Reference*, New York: Walter de Gruyter.
Schiffer, S. (1972) *Meaning*, Oxford: Clarendon Press.
Searle, J. R. (1969) *Speech Acts*, Cambridge: Cambridge University Press.
Sperber, D. and Wilson, D. (1986) *Relevance*, Oxford: Blackwell Publishers.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

21

G. H. von Wright (1916–)

FREDERICK STOUTLAND

Georg Henrik von Wright was born and educated in Helsinki, Finland, where his graduate work was supervised by Eino Kaila, a distinguished Finnish philosopher who was connected with the Vienna Circle, and who encouraged von Wright in that direction. Because of conditions on the continent, however, von Wright went to Cambridge in 1939 to study with C. D. Broad, and there he met Wittgenstein and Moore, who, together with Kaila, became the major influences on his philosophical development. He returned to Helsinki after a year and in 1941 received a Ph.D. for his thesis "The Logical Problem of Induction." After the war, he returned to Cambridge as a lecturer and in 1948 was appointed to the Chair of Philosophy in Cambridge, which Wittgenstein and, before him, Moore had held. In 1951 he resigned to return to Helsinki and resume the professorship to which he had been appointed in 1946. In 1961 he was appointed to the Academy of Finland, which at the time comprised twelve persons, whose lifetime membership in the Academy freed them to do their own work.

He has continued to teach and lecture around the world, notably as Professor-at-Large at Cornell, in giving (among others) the Gifford, Tarner, Woodbridge, Tanner, and Leibniz Lectures, and in engaging philosophers from many countries in philosophical conversation of the highest quality. He is the subject of a volume in the "Library of Living Philosophers" (1989), which contains a bibliography of over 400 papers and books (several dozen more have appeared since), essays on his work by thirty-one philosophers, and a notably instructive and interesting autobiography.

Von Wright's philosophical career has been marked by his working through one set of issues, then leaving it to focus on another set, and so on, and I have ordered my discussion of his contributions to reflect those stages in his career. The order is by no means exact, however, because certain themes have persisted throughout his work and because he often returns to issues when he has new things to say.

Induction and probability

Issues related to induction and probability were the focus of his work from his studies for his Ph.D. until the late 1940s. The topic of his doctoral dissertation was the "canons of induction," which Bacon and Mill used to ascertain causes and effects, and which von Wright replaced with the more precise notions of necessary and sufficient condi-

tions in order to restate and re-evaluate the classical canons. This was extended, corrected, and related to probability theory in a number of papers and then in his *Treatise on Probability and Induction*, written in 1948 but not published until 1951. He sums up this work as follows:

> I have tried to show how the probabilifying effect of evidence on a given hypothesis is an isomorphic reflection in numerical terms of a process of eliminating members from a class of hypotheses initially competing with the given one. This eliminative procedure is the logical core of Mill's canons. My aim can thus be said to have been a unification of the two main branches of inductive logic: induction by elimination in the tradition of Bacon and Mill, and induction by confirmation in the tradition founded by the Cambridge logicians J. M. Keynes, C. D. Broad, and W. E. Johnson and later continued by Carnap and others. (1993: 114)

His publications from this period also include studies of the paradoxes of confirmation, a topic to which he has returned (see 1983b), as he has to induction and probability in writing articles on each for the 1959 edition of the *Encyclopedia Britannica*.

Philosophical logic

Von Wright's work on induction and probability was not intended as a contribution to mathematical theories like those of von Mises or Kolmogorov, but as a philosophical study which used formal methods as a way of understanding and improving the concepts we employ in evaluating certain kinds of empirical hypotheses. The same is true of his work in logic: it aimed not to be a contribution to mathematical logic in the style of Gödel, Tarski, or Church but to be a philosophy of logic and a philosophical logic (see TARSKI, CHURCH, GÖDEL). He took the task of logic generally to be "to describe and systematize the principles used in argumentation, inference, and proof," the aim of his own work being philosophical reflection on those principles and the concepts they involve. In the course of that work, a new aim emerged: to extend the application of logic, as it had developed since Frege, to subject matters that traditionally made no explicit use of logical symbols and methods.

The focus of this work was the concept of logical truth, which he began to investigate by considering how far the Tractarian notion of logical truth as tautological could be extended to quantificational logic. He showed that this could be done for simpler quantificational structures, and in so doing invented "distributive normal forms," which others appropriated for technical uses (which included showing precisely how far the notion of a tautology could be extended). This work led to such related topics as conditionals, entailment, negation, and the logical antinomies, on which von Wright wrote clarifying and stimulating papers of the highest quality, from which much may be learned. (The earlier work is in *Logical Studies* (1957) and the more recent in "Philosophical Logic" (1983b).)

While working on the quantifiers, he noticed a parallelism between the structure of "some," "none," and "all" and that of the modal terms "possible," "impossible," and "necessary." Just as the negation of "Some *S* are *P*" is equivalent to "No *S* are *P*," so the negation of "*P* is possible" is equivalent to "*P* is impossible," and just as the negation of

"Some *S* are not *P*" is equivalent to "All *S* are *P*," so the negation of "Not-*P* is possible" is equivalent to "*P* is necessary," and so on. A little later, he noticed the same parallelism for the deontic modal terms "permissible," "forbidden," and "obligatory": "*A* is not permissible" is equivalent to "*A* is forbidden"; "Not-*A* is not permissible" is equivalent to "*A* is obligatory," etc. His project was to articulate these modalities in logical systems, analogous to propositional and predicate logic, which would have precise rules for well-formed formulae and valid inferences, which would enable exact determination of what a claim entailed and what it contradicted, and which would permit investigation of metalogical issues like consistency and completeness. His ideas on how to do this for the strict modalities were written in 1950 (published 1951b), while his proposals for the second appeared in his paper, "Deontic Logic," in the first issue of *Mind* for 1951 (reprinted 1957).

Von Wright showed that further concepts exhibited a similar structure, for example, the epistemic modalities "undecided," "falsified," and "verified" (if *P* is undecided, then *P* is not falsified; if *P* is falsified, then not-P is verified, etc.), and even time and causality can be seen as having a kind of modal structure. Von Wright worked on all these concepts, showing how logical principles, concepts, and methods could be applied to them, thus extending logical investigations into new areas. (For his summary, see 1993 (essays VI and VII) and 1989.) This inspired many others to apply logical techniques and symbolism to diverse philosophical topics, and while this work is too often technically or philosophically uninteresting, some of it has been very significant for both logic and philosophy. The latter, significant, work has prompted the view that this is the best – even the only – way to do analytic philosophy, but that is a view von Wright emphatically rejects.

Ethics, norms, and values

"Deontic Logic" is von Wright's most famous paper, for it created a new subject, which was his own in a special sense. It showed how to symbolize various normative claims and precisely determine their logical interrelations, and it opened up new claims and relations which are difficult to notice outside a logical system. This "logic of norms" remains of great interest to legal philosophers, for instance, who wrestle with such questions as whether two laws are mutually consistent or whether, and in what sense, a system of laws may be consistent or complete.

Von Wright's interest in the logic of norms was sustained because of its connection with issues about truth. He had since his youth been committed to what he calls a "deep conceptual gap separating the world of facts from that of norms and values," and he continues to hold that belief in its radical form as the view that normative judgments are neither true nor false. This raises the problem of how a logic of norms is even possible, for such a logic assumes that there are logical relations between norms and that there are disjunctive and conditional norm sentences, and those assumptions appear to require that norm sentences have truth-value.

In dealing with this problem, von Wright appeals to the distinction between sentences which *give* a norm, and thus are *prescriptive*, and those which *state* that a norm has been given, and thus are *descriptive*, which yields two possibilities for constructing a logic of norms without assuming that normative judgments are true or false. The first

is to have a logic, not of norms as such, but of the sentences which describe given norms and thus have truth-values. The second is to have a logic of norms proper but do not require that logically complex sentences with logical interrelations must have a truth-value, so that "Logic has a wider reach than Truth." Von Wright has changed his mind more than once on which of these ways is better (see 1963a, 1968, 1984), and more recently has suggested a third possibility, which is that deontic logic is "neither a logic of norms nor a logic of norm-propositions but a study of conditions which must be satisfied in rational norm-giving activity" (1993: 111; cf. 1996).

Norm and Action (1963a) and *An Essay in Logic and General Theory of Action* (1968) extend these logical investigations to the actions which norms govern and to the changes involved in actions (which I discuss below). But the study of norms is also part of moral philosophy, which von Wright was pursuing at the same time, largely because of his teaching duties; it resulted in *The Varieties of Goodness*, his main work in moral philosophy. Its central claim is that *moral* senses of "good," and of "right" or "duty," are derivative on non-moral uses of the terms. It is a conceptual inquiry, which aims at making "fixed and sharp that which ordinary usage leaves loose and undetermined," but since it denies that there is any clear distinction between ethics and metaethics, its inquiries into how to improve our moral concepts are also meant to be inquiries into what moral point of view we ought to adopt. It assumes that there can be "a philosophical pursuit deserving the name 'ethics', which shares with a common conception of 'meta-ethics' the feature of being a *conceptual investigation* and with a common concept of 'normative ethics' the feature of aiming at *directing our lives*" (1963b: 6). The book is Aristotelean in its insightful classification of the varieties of goodness, in its taking seriously the notion of virtue, in its taking the human good to be more basic than duty, and in its commitment to there being such a thing as *practical* rationality, which is not inferior to the rationality of the exact sciences. These notions are not rare today, but they were audacious in 1963, and it is understandable that von Wright regards *The Varieties of Goodness* as the best argued of all his works and the most fun to write.

Philosophy of action

Von Wright's work in philosophy of action stemmed from his attempt to get a symbolism adequate to distinguish the various kinds of action governed by norms: to distinguish, for example, what agents do (either by bringing something about or preventing something) from what they do not do (either intentionally or by simply doing nothing). It struck him that such distinctions could not be expressed without a symbolism for different kinds of *change*, and hence he worked at embedding a logic of action in a logic of change, a project to which he has returned a number of times (1963a, 1968, 1973).

Broader issues about action were implicit in this work and became explicit in *Explanation and Understanding* (1971). It argued that explanations of intentional action are logically distinct from explanations in the natural sciences because the latter appeal to causal laws while the former invoke conceptual connections between an agent's reasons and his actions. Explanation of action is, therefore, necessarily connected with practical reason (see 1983a), and the intentional attitudes that function as reasons should be seen, not as internal states with causal powers, but as ways of under-

standing and articulating what agents *mean* by their behavior. He connected this view with the Aristotelean tradition, contrasting it with the Galilean tradition which assimilated explanation of action (and of historical events) to the law-based model of the natural sciences.

Explanation and Understanding reinforced the rejection of a causal model of action by arguing that the concept of cause is inseparable from the concept of (experimental) intervention in nature, and hence inseparable from, indeed derivative on, concepts of action and agency. This implied that the kind of determinism which threatens to undermine the possibility of genuine human agency is self-defeating, a point von Wright has articulated in a number of ways (1973). Indeed, determinism has been a core interest of his since *Explanation and Understanding*, which has been expressed partly in logical investigations into issues such as kinds of necessity and knowledge and truth about the future (1984), partly in discussions of explanation and free will (1973), partly in criticisms of the way determinism encourages a "reified conception" of people and social institutions as governed by universal laws (1983a, 1993).

Philosophy of mind

Von Wright first encountered philosophy as a youth through the mind–body problem, but he wrote nothing about it until his Tanner Lectures, "Of Human Freedom" in 1984 (reprinted in *In the Shadow of Descartes*, 1998), when reflection on the relation of neuro-physiological explanation and action explanation led to some remarks on psycho-physical parallelism. He then began to write extensively on the philosophy of mind but published little until his 1998 book, which has papers on the relation of physiological and action explanation, on the concepts of quality and thing, and on perception and sensation (with special attention to sounds). The papers do not form a unity and do not discuss current literature and controversies (of which von Wright is by no means ignorant), but they manifest an intense effort to get clear on some of the most difficult and fundamental issues in philosophy of mind, and they contain a wealth of distinctions and observations which may prove productive.

Wittgenstein

Von Wright knew Wittgenstein well, was named as one of his literary executors, and has devoted enormous time and effort to his large and extraordinarily complex *Nachlass*. He has searched for lost material, interviewed persons who knew Wittgenstein, organized and indexed the papers (see *Wittgenstein*, von Wright 1982), and edited several volumes of his correspondence and manuscripts. While he has written relatively few papers on Wittgenstein's thought, these volumes include some splendid studies, which have been published (along with his moving "Biographical Sketch") in his 1982 book.

While von Wright was deeply impressed by his encounters with Wittgenstein, who no doubt influenced his work in numerous ways, he has chosen his own path. His work on philosophy of action and of mind are the most Wittgensteinian in content (his logical work is the least), but none of this can be traced back in any direct way to Wittgenstein's writing. All of his work, however, manifests Wittgenstein's stress on con-

ceptual multiplicity and his distrust of simple answers and philosophical theses, and it shares Wittgenstein's aversion to superficial, careless, or pretentious writing (see WITTGENSTEIN).

Humanism

Von Wright also shares with Wittgenstein a sensitivity to the larger context of philosophy, whose present character he also tends to think of as "the darkness of this time." He has written numerous essays (mostly in Swedish or Finnish) on writers such as Spengler, Toynbee, Dostoevsky, and Tolstoy and on topics including education, the state of the humanities, the image of science, the myth of progress, and the idea of revolution: essays aimed at clarifying the meaning of his life as a human being rather than as a professional philosopher. (See *The Tree of Knowledge*, 1993.)

He began as a believer in the "spirit of scientific rationality," which was inspired above all by post-Cartesian mathematics and physics, and which in turn inspired the philosophy of Kaila and the logical empiricists, the new logic, and analytic philosophy. He was never a believer in the idea that scientific rationality would result in inevitable progress – even his youthful "aesthetic humanism" was strongly affected by Spengler – but he was a believer in the ideal of philosophy as one of the exact sciences.

This attitude changed when he began teaching moral philosophy, and under the influence of Jaeger's *Paidea* he abandoned "aesthetic humanism" for an individualistic "ethical humanism," which recognized a practical rationality on a par with the rationality of the exact sciences. This was shattered by the effect the Vietnam war had on him, a war he protested in eloquent and effective ways and which forced him to think about the human condition in social and political terms. The result was the "social humanism" to which he remains committed and which he has expressed in powerful essays on the threat to human life of the very scientific rationality which inspired his early efforts in philosophy, essays which have made him Scandinavia's most prominent – if controversial – public intellectual.

His critical attitude toward the ideal of scientific rationality, above all for the major role it has played in creating social and technological conditions which both threaten the natural environment and inspire irrational and unjust ways to try to meet the threat, have also changed his conception of philosophy. He thinks of it, not as an exact science, but as critical reflection on the underlying assumptions of our thinking, judging, and acting – reflection which, since these assumptions are embedded in social institutions and traditions, must inevitably scrutinize the foundations of society (see "Intellectual Autobiography," 1989).

Although von Wright thinks of his humanist essays as parallel to rather than part of his philosophical work, they exhibit many of the same virtues. They are rich with illuminating insights, unexpected observations, and stimulating suggestions for further thought, and they are written in a style perfectly suited to their subject and to their author. His strictly philosophical work adds such further virtues as exact and intricate argument, systematic distinctions, and technical sophistication. These are the virtues prized by analytic philosophy, which von Wright has not abandoned, in spite of rejecting scientific rationality as the ideal of philosophy and being skeptical about its role as an ideal for any inquiry. What he regards as the chief virtue of philosophy has been

best expressed in the epigram to his autobiography, which is from Melville's *Moby Dick*: "All deep, earnest thinking is but the intrepid effort of the soul to keep the open independence of her sea."

Bibliography of von Wright's work

1941: *The Logical Problem of Induction*, Helsinki: Acta Philosophica Fennica (2nd revised edn., Oxford: Blackwell Publishers, 1957).
1951a: *A Treatise on Induction and Probability*, London: Routledge and Kegan Paul.
1951b: *An Essay in Modal Logic*, Amsterdam: North Holland.
1957: *Logical Studies*, London: Routledge and Kegan Paul.
1963a: *Norm and Action: A Logical Inquiry*, London: Routledge and Kegan Paul.
1963b: *The Varieties of Goodness*, London: Routledge and Kegan Paul.
1968: *An Essay in Deontic Logic and the General Theory of Action*, Amsterdam: North Holland.
1971: *Explanation and Understanding*, London: Routledge and Kegan Paul.
1973: *Causality and Determinism*, New York: Columbia University Press.
1981: *Freedom and Determination*, Helsinki: Acta Philosophica Fennica.
1982: *Wittgenstein*, Oxford: Blackwell Publishers.
1983a: "Practical Reason," *Philosophical Papers*, vol. 1, Oxford: Blackwell Publishers.
1983b: "Philosophical Logic," *Philosophical Papers*, vol. 2, Oxford: Blackwell Publishers.
1984: "Truth, Knowledge, and Modality," *Philosophical Papers*, vol. 3, Oxford: Blackwell Publishers.
1989: "Intellectual Autobiography," in *The Philosophy of Georg Henrik von Wright*, The Library of Living Philosophers, La Salle, IL: Open Court.
1993: *The Tree of Knowledge and Other Essays*, Leiden: Brill.
1996: *Six Essays in Philosophical Logic*, Helsinki: Acta Philosophica Fennica.
1998: *In the Shadow of Descartes: Essays in the Philosophy of Mind*, Dordrecht: Kluwer Academic Publishers.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

22

Roderick Chisholm (1916–1999)

Part I: Epistemology

RICHARD FOLEY

Part II: Metaphysics

DEAN ZIMMERMAN

Roderick Chisholm's work spans six decades and an impressive range of subjects. His books and articles on Brentano and Meinong, together with his work as a translator and editor, and as director of the Brentano Foundation, brought Anglo-American analytic philosophy back into contact with the riches of the Austrian philosophical tradition. He wrote several important papers on the foundations of ethics and axiology (e.g. Chisholm 1963, 1974). But Chisholm is best known for his many contributions to epistemology and metaphysics.

Part I: Epistemology

The most important of Roderick Chisholm's writings on epistemology are *Perceiving*, *The Foundations of Knowing*, and the three editions of *Theory of Knowledge*. In these and in his other works, Chisholm addressed virtually every major problem in epistemology. At the heart of his epistemological system is a set of epistemic principles that are intended to generate intuitively plausible results about the degree to which various propositions are justified for an individual. The key to Chisholm's epistemology is understanding how these principles fit together and also understanding both their epistemological status (how is it that we can come to know them?) and their metaphysical status (are they necessary or contingent, and what is it that makes them true?).

Terms of epistemic appraisal

In formulating his epistemological principles, Chisholm presents a set of terms of epistemic appraisal, which he defines using a basic, prephilosophical notion of justification. The following are simplified versions of the definitions that appear in the third edition of *Theory of Knowledge* (1989a).

Certain: A proposition *p* is certain for an individual *S* if and only if no other proposition is more justified for *S* to believe.

Evident: A proposition *p* is evident for *S* if and only if *S* is at least as justified in believing *p* as withholding judgment on that which is counterbalanced.

Beyond reasonable doubt: A proposition *p* is beyond reasonable doubt for *S* if and only if *S* is more justified in believing *p* than withholding judgment on *p*.

Epistemically in the clear: A proposition *p* is epistemically in the clear for *S* if and only if *S* is at least as justified in believing *p* as withholding judgment on *p*.

Probable: A proposition *p* is probable for *S* if and only if *S* is more justified in believing *p* than disbelieving *p*.

Counterbalanced: A proposition *p* is counterbalanced for *S* if and only if *S* is as justified in believing *p* as believing not-*p*, and vice-versa.

Chisholm intends the first five of these terms to be such that the higher ones imply the lower ones, and he introduces axioms to ensure this (1989a: 12, 13, 17). So, if a proposition is certain for someone, it is also evident for that person, and if it is evident for the person, it is also beyond reasonable doubt, and so on down the list.

Epistemic principles

Making use of the above terms of epistemic appraisal, Chisholm proposes a set of epistemic principles. The principles are expressed as conditionals, whose antecedents describe sufficient logical conditions for the application of these terms of epistemic appraisal. In the most straightforward case, a principle will assert that if certain non-epistemic conditions are satisfied (e.g. conditions about what someone is experiencing, believing, etc.), then a proposition *p* has a certain epistemic status for the person (e.g. it is evident or beyond reasonable doubt).

Chisholm's project in formulating these principles can be compared to a traditional project in ethics. A central aim of theoretical ethics is that of describing a set of non-moral conditions that is sufficient to make an action morally right. According to utilitarians, the non-moral conditions are ones having to do with the production of pleasure and the avoidance of pain. If of all the alternatives available to me, alternative *X* will produce the greatest balance of pleasures over pains, then I am required to do *X*. Utilitarians claim that this is *the* fundamental principle of morality. For them, there is but one source of moral obligation. Others disagree, insisting that there are other sources as well, ones that are not directly concerned with the maximization of happiness. Equality and fairness are among the usual candidates. If doing *X* would produce a fair result, then, according to this view, I have a prima-facie obligation to do *X* even if doing so would not maximize happiness. There may be other sources as well, and corresponding to each of these sources will be an ethical principle, asserting that the source in question produces a prima-facie moral obligation.

This latter view is the counterpart of Chisholm's view in epistemology. He thinks that there is more than one source of epistemic justification, and corresponding to each of these sources is an epistemic principle describing the conditions under which the source produces justification. However, Chisholm believes that some of these sources produce justification only in conjunction with other sources. Thus, the epis-

temic principles corresponding to these sources must make reference to the workings of other principles. The result is a collection of principles that are interdependent in complex ways.

Below are some of the most important of the epistemic principles that Chisholm defends:

1 If *F* is a self-presenting property and if *S* has *F* and if *S* believes himself (herself) to have *F*, then it is certain for *S* that he (she) has *F*.
2 If it is evident to *S* that he (she) is appeared to ϕ-ly and it is epistemically in the clear for *S* that something is appearing to him (her) in this way, then it is evident for *S* that something is appearing ϕ to him (her).
3 If it is evident to *S* that *S* is appeared to ϕ-ly and if *S* believes that it is a *G* that is appearing to him (her) in this way and if this proposition is epistemically in the clear for S, then it is beyond reasonable doubt for *S* that he (she) perceives a *G*.
4 If *S* believes a proposition that is not disconfirmed by the set of propositions that are evident for *S*, then the proposition is probable for *S*. (According to Chisholm, *p* disconfirms *q* amounts to *p* tends to make not-*q* probable.)
5 If *S* believes a proposition that is not disconfirmed by that which is probable for *S*, then the proposition is epistemically in the clear for *S*.
6 If there are three or more concurrent propositions and if each of them is epistemically clear for *S* and if in addition one of them is beyond reasonable doubt for *S*, then they all are beyond reasonable doubt for *S*.
7 If there are three or more concurrent propositions and if each of them is beyond reasonable for doubt for *S* and if in addition one is evident for *S*, then they are all evident for *S*.

Principle (1) makes reference to "self-presenting properties," which Chisholm takes to be purely psychological properties. These properties are non-relational, in the sense that from the fact that I have a property of this sort, nothing logically follows about how I am related to the non-psychological world. For example, from the fact that I have the property of thinking about sailing my boat, it does not follow that I am in fact sailing my boat. I does not even follow that I have a boat. Nor, says Chisholm, does anything else follow about the non-psychological world. On other hand, from the fact that I have the property of being stuck in a traffic jam and thinking about sailing my boat, something does follow about the non-psychological world. It follows that there are traffic jams, that I am in one, and so on. So, this property is not a self-presenting one.

Chisholm distinguishes two kinds of self-presenting properties: intentional properties (ways of thinking, hoping, fearing, wondering, wishing, desiring, intending, etc.) and sensible properties (ways of being appeared to by the various senses). Principle (1) says that if I have a self-presenting property and if I believe that I have it, then the proposition that I have the property is maximally justified for me. Nothing is more justified for me to believe.

The self-presenting provides a foundation on the basis of which other contingent propositions can come to have justification. Principle (2) describes one way of this happening. If it is evident to me that I am having a visual experience of the sort that is involved in seeing a cat and if in addition it is epistemically in the clear for me that something is appearing to me in this way, these two things combine to make it evident

for me that something is appearing to me in this way. It may not be evident to me whether it is a cat or a dog or a bush that is appearing to me, but it is evident for me that something is doing so. It is evident, in other words, that I am not hallucinating.

Chisholm also proposes principles of "perceptual taking." For example, the above principle (3) implies that if it is evident to me that I am appeared to in a certain way and if I believe that it is a cat that is appearing to me in this way and if moreover this proposition is epistemically in the clear for me, then these three things combine to make it beyond reasonable doubt for me that I perceive a cat. There is also a principle analogous to (3) for memory, expressed in terms of what I seem to remember (1989a: 68).

The antecedents of principles (2) and (3) make reference both to propositions that are evident and propositions that are being epistemically in the clear. Principle (1) describes how propositions can become evident for me, but Chisholm believes that the set of propositions that are epistemically in the clear is much larger than the set of evident propositions. Thus, there must be some other source of epistemic justification for them. What is this other source?

Chisholm says that it is belief itself, that one way in which a proposition can obtain a degree of epistemic justification is by being believed. Principles (4) and (5) are meant to describe how. According to (4), if I believe a proposition that is not disconfirmed by the set of propositions that are evident for me, then the proposition is probable for me. A large number of propositions can become probable for me in this way. They will have this weakly favorable epistemic status even if there is no other positive source of justification for them – from self-presentation, perception, or memory, for example. Moreover, principle (5) allows these propositions to rise to an even higher epistemic status. According to (5), if I believe a proposition that is not disconfirmed by the set of other propositions that are probable for me, then this proposition is epistemically in the clear for me. What are these propositions that are at least probable for me? In large part, they are propositions that satisfy the antecedent of principle (4), namely, believed propositions that are not disconfirmed by that which is evident for me. So, (4), as it were, creates much of the material for (5) to do its work.

Principles (4) and (5) are principles of negative coherence. Together they imply that if a believed proposition is not incoherent with the set of other propositions that are probable for me (many of which get this status by the fact that I believe them and they are not disconfirmed by that which is evident for me), then it is acceptable for me to believe the proposition.

With these principles in hand, reconsider the question of how the propositions mentioned in the antecedent of these principles (2) and (3) get the status of being epistemically in the clear for me. Principles (4) and (5) provide an answer. They can get this status by being believed by me. If I believe a proposition of the sort mentioned in the antecedent of (2), say the proposition that something is appearing to me in a cat-like way, and if the propositions that are probable for me do not disconfirm this proposition, then the proposition is epistemically in the clear for me. And then, principle (2) says that this in conjunction with the fact that it is evident to me that I am appeared to in a cat-like way makes it evident that something is appearing in a cat-like way to me. It is evident that I am not hallucinating.

Similarly for propositions of the sort mentioned in the antecedent of (3): if I believe that it is a real cat appearing to me in a cat-like way and if this proposition is not dis-

confirmed by the set of other propositions that are probable for me, then the proposition is epistemically in the clear for me. And then, principle (3) says that this in conjunction with the fact that it is evident for me that I am appeared to in a cat-like way makes it beyond reasonable doubt that it is a cat – and not, say, a dog or a bush – that I am perceiving.

Chisholm also thinks that relations of positive coherence among a set of propositions, or what he calls "concurrence relations," are an important source of justification. A set of propositions is concurrent just if the propositions are logically independent and mutually supportive, in the sense that each proposition in the set is such that the others tend to make it probable.

Chisholm defends two principles of concurrence. Principle (6) says that if there is a set of concurrent propositions each of which is epistemically in the clear for me and at least one of which is also beyond reasonable for doubt for me, then they all become beyond reasonable doubt for me. Principle (7) says something similar for concurrent propositions of the next highest epistemic status. According to (7), if there is a concurrent set of propositions each of which is beyond reasonable for doubt for me and at least one of which is evident for me, then they all become evident for me.

So, despite his reputation as the leading foundationalist, Chisholm is also a coherentist. But unlike a pure coherentist, he does not think that positive coherence relations are the only source of empirical justification.

Together, the above principles describe what Chisholm takes to be some of the principal sources of empirical justification: namely, self-presentation, perception, memory, belief coupled with a lack of negative coherence, and, finally, positive coherence among propositions with some antecedent positive epistemic status.

The epistemological and metaphysical status of the principles

According to Chisholm, we have at least a vague, prephilosophical idea of what it is for a belief to be justified (1989a: 5), an idea which guides us in identifying instances of beliefs that are justified. In turn, these intuitions about justified beliefs allow the epistemological project to get off the ground. Chisholm is a particularist when it comes to matters of epistemological method (1989a: 7). He begins by examining particular instances of beliefs that he takes to be justified, and he then tries to abstract out of these instances general conditions of justification, which he expresses in the form of epistemic principles.

Chisholm also presupposes that we can improve and correct our beliefs by reflection, eliminating those that are unjustified and adding others that are justified (1989a: 1, 5). This presupposition acts as a constraint when he tries to use particular instances of justified belief to formulate general conditions of justification. It forces him to look for conditions to which we have reflective access, since otherwise there would be no reason to think that we could eliminate unjustified beliefs and add justified ones simply by being reflective. This is one of the senses in which Chisholm is an internalist about justification, in an epistemic sense.

The prephilosophical notion of justification that allows epistemology to get off the ground is vague, like most ordinary notions, but it need not remain so. One of the beneficial by-products of formulating and refining epistemic principles is that the basic

notion becomes increasingly precise, so that eventually epistemolgists are in a position to give a general characterization of it. According to Chisholm, the characterization is to be given in ethical terms. Epistemic justification is ultimately to be understood in terms of ethical requirements on our believings and withholdings. More specifically, to say that an individual *S* is more justified in believing *p* than withholding on *p* is to say that *S* is required to prefer the former over the latter (1989a: 59). Chisholm goes on to claim that requirements to prefer are best explicated in a negative way. The requirement to prefer believing *p* over withholding on *p* is a requirement not to choose between believing and withholding without choosing the former, and this, he points out, is a requirement that can be satisfied even if one does not have direct control over one's believings and withholdings.

In addition, Chisholm says that this requirement is one that supervenes on non-normative states, specifically, on conscious states (1989a: 60). As such, a proposition could not have an epistemic status different from the one it does have for an individual without that individual's psychological states being different. Thus Chisholm takes his epistemic principles to express necessary truths, and the truths that they express are ultimately ones about the relationship between an individual's conscious psychological states at a time and an ethical requirement on believings and withholdings.

This illustrates another sense in which Chisholm is an internalist about justification, a metaphysical sense. The conditions that make a proposition evident or beyond reasonable doubt or probable are internal conditions. They are current, psychological states, not non-psychological "external" states, and not past psychological states. Chisholm's epistemic internalism requires something in addition to this. It requires that we always have reflective access to these internal conditions.

The definition of knowledge

The epistemic principles and the terms of epistemic appraisal used to formulate the principles constitute the heart of Chisholm's epistemological system. They are the tools Chisholm uses to address the major questions of epistemology. Among these questions, none has preoccupied Chisholm more than the question, What is knowledge?

Over his career, he proposed various definitions of knowledge, most of them variants of the idea that knowledge is non-defectively evident true belief. Like many proposed definitions of knowledge, Chisholm's definition was aimed at coming to grips with a pair of examples presented by Edmund Gettier, which were designed to illustrate that knowledge cannot be adequately defined as justified true belief. The basic idea behind both counterexamples is that one could be justified in believing a falsehood P, from which one deduces a truth Q. In this case one has a justified true belief in Q but does not know Q. Gettier's examples inspired a host of similar counterexamples, and the search was on for a fourth condition of knowledge, one that could be added to belief, truth, and justification to produce an adequate analysis of knowledge.

The two most distinctive aspects of Chisholm's attempt to handle Gettier problems are, first, his insistence that a belief must be evident to count as knowledge, and second, his insistence that what makes the belief evident must be non-defective (Chisholm adds some further qualification; see 1989a: 98).

"Evident" is among the strongest of Chisholm's terms of epistemic appraisal, ranking only below that which is certain. A proposition is certain for an individual *S* only if it is maximally justified; no other proposition is more justified for *S* to believe. Among the propositions that can be certain are simple necessary truths, for example, the elementary truths of arithmetic, as well as contingent propositions about self-presenting states. In defining knowledge in terms of the evident, Chisholm is rejecting the view that knowledge requires certainty. On the other hand, he is insisting that knowledge involves a very high degree of justification. For Chisholm, a paradigmatic requirement on believing is that of withholding judgment on that which is counterbalanced, for example, the proposition that the next toss of a fair coin will turn up heads. A proposition is evident for *S*, in turn, only if *S* is at least as justified in believing it as withholding judgment on that which is counterbalanced. For Chisholm, this represents a very high degree of justification.

Moreover, if *S* is to have knowledge of a proposition, not only does the proposition have to be evident for *S*, in addition that which provides this very high degree of justification must be non-defective, in the sense that it must not make any falsehood evident for *S*. To illustrate the intuitive force of this requirement, consider one of Gettier's examples. Smith has very strong evidence for the proposition that Jones owns a Ford, since Smith is aware that Jones has always owned a car, that the car has always been a Ford, that Jones has just offered Smith a ride while driving a Ford, and so on. From this evidence and the proposition that Jones owns a Ford, Smith deduces the disjunctive proposition, either Jones owns a Ford or Brown is in Barcelona. He infers this proposition despite having no idea of where Brown is. However, it turns out that Jones does not in fact own a Ford (he has been driving someone else's car) while, by chance, Brown is in Barcelona. So Smith has very strong justification for the proposition that either Jones owns a Ford or Brown is in Barcelona; he believes the proposition; and the proposition is true. But it seems as if Smith does not know the proposition. Why? Chisholm's answer is that although the proposition may be both true and evident for Smith, the considerations which makes the proposition evident for him also makes evident a falsehood, namely, that Jones owns a Ford.

Part II: Metaphysics

Whereas Chisholm's epistemological views constitute a unified whole that may be usefully and concisely summarized, the many metaphysical problems he addressed form a heterogeneous collection that does not submit readily to concise overview. Furthermore, his role in the rejuvenation of metaphysics during the second half of the twentieth century would not be conveyed by a summary of his approaches to particular problems. One must back up a bit to see why and where his influence was great.

Chisholm's impact on contemporary metaphysics would be hard to overestimate. By the end of 1950s his contributions were already numerous. He mounted an influential defense of the meaningfulness of traditional metaphysical questions in the face of deflationary critiques from the "ordinary language" philosophers (1951, 1952, 1964). He helped bring down the curtain on phenomenalism (1948, 1957b: Appendix). He drew attention to Franz Brentano's characterization of the psychological in terms of "intentional inexistence," and attempted to rehabilitate it as a logico-linguistic criterion of

sentences reporting intentional mental states (1955–6). He defended the thesis that linguistic intentionality is to be explicated in terms of the intentionality of thought, and not the reverse (1955a, 1957a). And he helped focus debates about counterfactuals, dispositions, and laws of nature (1946, 1955b). This work was widely anthologized in subsequent decades.

But his most important contributions to metaphysics came somewhat later, as he began to construct complex, evolving, and interconnected theories of action, persistence through time, events and causation, reference and intentionality, and ontological categories.

A summary of the positions he defended on these topics would go some way toward explaining his importance as a metaphysician. In many areas, one still finds Chisholm's work cited as containing the paradigmatic formulation of an important position, or the original statement of a paradox that stands in need of resolution. In action theory, for instance, his defense of the incompatibility of freedom and determinism, and of agent causation, are as frequently discussed as ever. But a simple survey of Chisholm's views on particular metaphysical issues would miss the forest for the trees. A better picture of his place in twentieth-century metaphysics can be gained by considering the status of metaphysics at the time his career began, and by comparing Chisholm's methodology with that of another imposing figure from the same generation: W. V. Quine.

Metaphysics at mid-century

Most philosophical work that bears the (sometimes pejorative) label "metaphysics" is characterized by its attention to matters of ontology. A central part of the discipline has always been the construction of comprehensive ontological schemes, theories about the nature of and relations among the most abstract categories under which absolutely *everything* falls, together with the explicit use of these ontological distinctions in the formulation of solutions to philosophical problems. Indeed, one could argue that distinctively metaphysical problems always involve the very abstract categories appropriate to ontology; and that any philosophical problem becomes, at least in part, a metaphysical problem as soon as ontological distinctions become central to its statement and resolution.

By mid-century, metaphysics of this sort had fallen on hard times. The air had gone out of debates about the ontological status of universals and particulars, the distinction between essence and accident, and so on. Russell and Moore and, perhaps, the tractarian Wittgenstein had taken such questions pretty seriously (see MOORE, RUSSELL, and WITTGENSTEIN); but logical empiricism, Wittgensteinian "therapy," and Austinian "ordinary language philosophy" had eclipsed the metaphysical preoccupations of the earliest "analytic philosophers" (see AUSTIN and WITTGENSTEIN). The proponents of these influential doctrines all thought (albeit for different reasons) that the traditional questions of metaphysics were misguided, unanswerable, nonsensical.

Further, the reputation of metaphysics was poorly served by the obscurity of many of its most well-known practitioners. Clarity of exposition was not among the virtues exemplified by Royce, Bradley, Bosanquet, Bergson, Whitehead – names that *meant* metaphysics at the time. To the skeptical, it could easily seem that the recipe for success in metaphysics was this: (1) invent your own baroque ontological scheme, using a new,

peculiar jargon; (2) claim that it is radically opposed to all preceding metaphysical systems; and (3) explain its intricacies by the introduction of further undefined technical terms in a series of ever longer books. Chisholm's chief contribution to contemporary metaphysics was to show, by precept and, more importantly, by example that it is possible to construct metaphysical systems on a grand scale without falling into these vices. He championed a chastened approach to metaphysics, one that neither shies away from the traditional problems of ontology, nor falls back into the arcane, untethered system-building that had given metaphysics a bad name.

The comparison with Quine

W. V. Quine began teaching at Harvard while Chisholm was a graduate student. Quine provided something that would prove crucial to Chisholm's metaphysical program: the approach to questions of ontological commitment defended in "On What There Is" (1948), but already in place by 1939, when Chisholm was a student (Quine 1939). Chisholm took Quine's criterion of ontological commitment to amount to the following injunction: If one affirms a statement using a name or other singular term, or an initial phrase of "existential quantification," like "There are some so-and-so's" (see QUINE), then one must either (1) admit that one is committed to the existence of things answering to the singular term or satisfying the description, or (2) provide a "paraphrase" of the statement that eschews singular terms and quantification over so-and-so's. Both Quine and Chisholm agree that Meinong, who affirms truths about all sorts of things which he then admits do not exist, is trying to have his cake and eat it too; Meinong must be resisted if metaphysics is to be kept honest.

Chisholm's metaphysics looks nothing like Quine's, however. For Quine, it is the deliverances of science alone that need be taken into account when attempting to work out one's ontological commitments; he identifies the project of "limning the true and ultimate structure of reality" with that of working out the most ontologically austere regimentation of the language of the harder sciences. This enables Quine to keep his ontology lean, including nothing but the most well-understood, sharply demarcated things: ultimately, nothing but concrete spatiotemporal entities and the abstract but well-defined world of set theory. But the cost is great: the repudiation of quite a lot of what we would ordinarily regard as truisms about beliefs, desires, and other intentional attitudes; about what must or might be the case; about what would have happened if . . . ; and so on (see, e.g., Quine 1960). Chisholm, however, asks: Why not assume, in the seminar room, the same things we take ourselves to know in everyday life? Why, when we do philosophy, should we appeal to nothing but what we find in our physics and chemistry textbooks? Chisholm rejects Quine's skepticism toward all but science; an ontological scheme must show its adequacy on a much broader playing field.

Both Chisholm and Quine agree that ontological schemes are to be judged by the competing desiderata of simplicity and sufficiency of scope. One scheme is simpler than another if it posits fewer, and better understood, types of entities. One scheme is superior to another in scope insofar as it allows for the statement of satisfactory philosophical theories on more subjects, theories that preserve, sometimes in the face of apparent contradiction or philosophical puzzlement, most of what we take ourselves to know.

Quine's austere ontological naturalism is purchased at the cost of severe restrictions on the scope of what we may reasonably take ourselves to know. Although one cannot accept the mathematics needed for science without set theory, no further "queer entities" need be recognized by one who affirms nothing but the deliverances of the (sufficiently hard) sciences. Chisholm, however, has many more truths to consider; for him, balancing the competing desiderata of simplicity of scheme and sufficiency of scope is much trickier. The adequacy of an ontological scheme comes to turn upon its role in the resolution of the traditional problems of philosophy, most of which Quine was able to sidestep by rejecting the commonsensical convictions from which the problems arise.

It is no surprise, then, to find the two philosophers differing drastically despite their initial point of agreement. Chisholm finds that one cannot arrive at metaphysical theories satisfying both desiderata of simplicity and scope without making reference to things not found in Quine's ontology, such as "intensional objects." He can find no ontologically perspicuous theory that does justice to what we know about persons while eschewing irreducibly intentional (psychological) notions (e.g. "conceiving," "attributing"). Ultimately, he concludes that persons must be very special indeed: they have causal powers unlike those found elsewhere in nature, they can "grasp" or conceive of abstract objects, and their persistence conditions are mysteriously different from those of ordinary physical objects. Quine, and many other naturalistically inclined philosophers, will find such conclusions fantastical. Be that as it may, the theories Chisholm constructs offer solutions to a host of philosophical problems; and his metaphysical program stands as a challenge to be met by those who would be more naturalistic or nominalistic than Chisholm, but who are not prepared to retreat into a skepticism as radical as Quine's.

Chisholmian methodology illustrated: states of affairs as necessary things

Chisholm's ontological views underwent frequent revision, as one or another scheme proved inadequate in scope, unable to make room for enough of what we take ourselves to know; or as he thought of some way to keep a plausible philosophical theory in place while simplifying its ontological commitments. One of the more radical changes was the rejection of "states of affairs" in the early 1980s. It provides a good example of Chisholm's effort to make systematic metaphysics responsible by tying ontology to the resolution of a wide spectrum of philosophical problems. In this case, the change was brought about by problems of self-reference.

The greatest ontological divide in Chisholm's theory of categories is between necessary things and contingent things. The states of affairs so central to Chisholm's ontology throughout the 1960s and 1970s were taken to be necessary things.

Chisholm advanced several ways of marking the distinction between necessary and contingent things. He hoped to restrict his modal primitives to those expressible by means of one locution: "*x* is necessarily such that it is *F*," where "*F*" can be replaced by any predicate, and the phrase is equivalent to "*x* is necessarily such that it exists if and only if it is *F*." (Something is possibly *F*, of course, if and only if it is not necessarily not *F*.) But then, even if "exists" were allowed as a predicate substitutable for *F*, replacing *F* with "exists" yields only a sense of "necessarily existing" according to

which *everything* exists necessarily. One proposal for making the distinction within his restricted vocabulary is this: contingent things are possibly such as to be coming into existence or passing away (i.e. possibly such as to have had no properties and possibly such as to be going to have no properties) and necessary things are neither (cf. 1989b: 164, and 1996: 127). This presupposes that there are no things that could have failed to exist but that, given that they do exist, cannot possibly be created or destroyed. Some might have doubts about this assumption. Chisholm may have doubted it himself, since he tried other ways of making the distinction.

Another proposed mark of the necessary/contingent divide is this: x is necessary if and only if x has a property that is essential to it and that nothing else could possibly have (i.e. a property that is an individual essence of x); and everything is such that something has that property (1986: 26). This presupposes that every necessary thing has an individual essence. Perhaps Chisholm had doubts about this assumption, too; for his last attempt to formulate a criterion for the necessary existence of a thing, x, was: "There is an attribute that is such that (1) everything is necessarily such that there is something having that attribute, (2) x is necessarily such that it has that attribute, and (3) that attribute is not necessarily had by everything" (1996: 17). Counterexamples are generated if one allows for disjunctive properties such as *being either an animal or prime*. But Chisholm also developed theories of the structure of properties, including accounts of what it is for a property to be a disjunction of two others. Perhaps he would have found the resources there to refine his last definition so as to rule out such counterexamples.

Chisholm long held that there were at least two sorts of necessary thing: states of affairs and properties or attributes. (He always at least left an opening in his table of the categories for a third, as well: God, a necessary substance upon which all else depends.) He advocates an "intentional approach" to both states of affairs and properties; that is, he claims that their criteria of identity and structural features can only be adequately described using intentional terms, such as "believing" and "conceiving." States of affairs are defined as those things which one may believe (1976: 117), properties as those things which one may believe to be exemplified by other things (1996: 29). Both are given intentional criteria of identity. A state of affairs p is identical with a state of affairs q if and only if, necessarily, (1) p "obtains" or "occurs" if and only if q does (1976: 118); and (2) whoever believes p believes q, and vice versa. A property F is identical with a property G if and only if, necessarily, (1) something exemplifies F if and only if it exemplifies G, and (2) whoever conceives F conceives G, and vice versa (1989b: 145). *Propositions* are identified with states of affairs that either always obtain or never obtain, *events* with obtaining states of affairs that are not propositions and that entail the exemplification of a certain sort of property – a property "rooted in" the time at which it is exemplified.

Chisholm's ontology is "realist" in several senses of the term. It includes properties that, like Plato's universals, exist whether or not they are exemplified. It is opposed to psychologism and linguisticism about the subject matter of logic. Logic discovers necessary relations among *propositions*: necessarily existing states of affairs, in no sense mind-dependent or language-dependent. Indeed, the true propositions are not to be distinguished from *facts*. And so Chisholm advocates what is sometimes called an "iden-

tity theory of truth": true propositions "correspond with facts in the fullest sense that is possible, for they *are* facts" (1977: 88).

The first person and the rejection of haecceities

Throughout the 1960s and 1970s states of affairs figure prominently in Chisholm's metaphysics, epistemology, and metaethics. Here are some examples of the many duties they perform. Statements about particular occurrences of a given type of event are to be paraphrased in terms of the "obtaining" of abstract, eternal states of affairs. And a causal relation between a pair of events is really a matter of two states of affairs being causally related *relative to a certain time*. On this approach, there is no need to recognize an ontological category of "tropes" or "particularized properties" in addition to states of affairs and properties conceived as Platonic universals (1976: ch. IV). "Times" are given a gloss much like A. N. Prior's: they are maximal, consistent states of affairs, complete ways the world could be "all at once" (1979a: 357). Belief and other propositional attitudes are said to be relations between thinkers and states of affairs (1976: ch. IV). More generally, a relatively simple ontology of properties, states of affairs, and contingent, persisting individual things appears to be adequate to the formulation of philosophical theories across the whole range of subjects Chisholm addressed in this period.

The phenomenon of first person reference subjected the ontology of states of affairs to considerable strain. If propositional attitudes are relations between thinkers and states of affairs, what states of affairs are implicated in those attitudes expressed using the first person pronoun? Ernst Mach catches sight of himself in a mirror without realizing who it is, and thinks: "That is a shabby pedagogue," without thinking: "*I* am a shabby pedagogue." The contents of the two attitudes differ; but how is this difference to be reflected as a difference in the structures of states of affairs, as it must be on Chisholm's theory? Since states of affairs are necessary things, their constituents, too, must be necessary existents. The only way, then, for a state of affairs to be *about* some contingent thing is for it to contain an "individual concept" of that thing: a property only one thing could have, and one that is had by that thing. But what individual concept is involved in my first person thoughts? What extra property is there in the state of affairs *I am a shabby pedagogue* that is not present in *Someone is a shabby pedagogue?* Surely I need not know anything about my relations to other things in order to think a first person thought; so it must be some intrinsic property, peculiar to me, that enables me to think of myself in this way. And so Chisholm is led to accept the notion that each person has an "haecceity," an individual essence peculiar to him or her, and "repugnant to" everything else (1976: ch. I).

Chisholm gradually came to feel that introducing haecceities for this purpose was a cheat. Although extraordinarily useful, haecceities remain, at bottom, utterly mysterious. We cannot rest content with simply positing their existence on the basis of their usefulness, since a part of their use is supposed to be their accessibility to intellectual grasp by the thinking things that exemplify them:

> If this essentialistic theory were true, then every time a person expresses himself by means of an I-sentence he grasps his own essence or haecceity. But, one wonders, do I *ever* thus grasp my own individual essence or haecceity? If I do ever grasp it, shouldn't I be able to

> single out its various marks? . . . [I]f I can grasp my individual essence, then I ought also to be able to single out in it those features that are unique to it. If *being me* is my individual essence and *being you* is yours, then, presumably, each analyses into personhood and something else as well – one something else in my case and another in yours. But I haven't the faintest idea what this something else might be. . . .
>
> I think that Brentano was right about this point. He said that, when we consider the nature of ourselves, we *never* grasp any properties that are individuating. Any property I know myself to have is one which is such that some entity other than I could also have that property. (1979a: 322)

The phenomenological inadequacy of the haecceity theory led Chisholm to rethink problems of self-reference, looking for an haecceity-free theory that would allow for the distinctions we actually make among self-directed beliefs. What resulted was the "direct attribution" theory of belief: the objects of the so-called propositional attitudes are really *properties*, and the things that are true and false (in at least one primary sense) are *direct attributions* of properties to oneself (Chisholm 1979b, 1981). (David Lewis reached the same conclusion independently: Lewis 1979.) Forced to regard the objects of belief, hope, wonderment, etc. as properties in at least those cases ascribable by means of an indirect reflexive ("she, herself," "he, himself"), Chisholm (and Lewis) advocate treating all believing, etc. as a matter of the self-ascription of properties. When a person believes that she, herself, is mortal, she self-ascribes the property of mortality. When she believes, with respect to her father, that he is mortal, what is happening is this: she self-ascribes a property that implies that there is some relation holding between her and only one other person, and that person is mortal; and her father in fact stands in that relation to her. When she believes that someone (or other) is mortal, she self-ascribes a simpler property: *being such that someone is mortal*.

In *The First Person*, Chisholm works out interpretations of demonstratives, of proper names, and of sense and reference, in terms of the self-ascription of properties. As in his correspondence with Sellars (Chisholm 1957a), he defends the primacy of psychological intentionality over linguistic intentionality: we conceive of and self-ascribe properties that allow us to single out other things (cf. SELLARS), and we use words to cause others to conceive of and self-ascribe properties that single out the same things. In Chisholm's view, there is no way to avoid positing an irreducible intentional relation, such as "conceiving," that relates thinkers to extramental things (properties and, indirectly, other individuals) – a relation that cannot be identified with an ability to manipulate words in either a public language or an inner "language of thought."

The unraveling of the ontology of states of affairs

At first, the new account of the propositional attitudes sent relatively minor ripples through Chisholm's system, as he examined the extent to which the change called for modifications of his views in epistemology (1982: ch. 1), action theory, axiology, and ethics, and of his resolution of the paradox of analysis (1986). In these areas, there was little change in fundamental doctrine. But a fairly radical rethinking of his theory of events and causation was called for. The self-ascription account of thinking solves problems with the older, propositional account by rejecting the received opinion that truth and falsity are, at bottom, properties of propositions. In order to give a unified

theory of truth and falsehood, Chisholm adopts what he calls a "doxastic theory of truth," not unlike Russell's "multiple relation theory of judgment" (Russell 1910): it is beliefs or judgments that are true and false in the "primary sense"; the truth and falsity of other things is to be explicated in terms of the sense in which beliefs are true and false (1993, 1986: 23). This strips the old states of affairs of two of their most important functions: as the things that are, at bottom, true and false; and as the objects of propositional attitudes. Furthermore, now that haecceities have been rejected, no necessarily existent state of affairs can, in any obvious way, imply the existence of contingent particulars; so states of affairs are inadequate vehicles of truth and falsehood in all but the most abstract or general cases. States of affairs become a third wheel within the theory of the true and false, and are eventually jettisoned.

But states of affairs had played a dual role, as both objects of propositional attitudes and, when true, worldly facts and events. Formerly, the bearers of truth and falsehood were propositions, which were a category of states of affairs; and states of affairs (when they obtained) were not to be distinguished from facts. When the bearers of truth and falsehood are doxastic – acts of judgment – no simple identification of the bearers of truth with facts or events is possible. Many facts and events have nothing to do with judgments or thinkers. Something must be introduced to play the roles of *fact* and *event* in the new ontology: those things in virtue of which acts of judgment are true or false, and the sorts of things that are causes and effects. And so Chisholm introduces a new category, that of *states*: contingently existing structures that are made out of things and properties, and that exist only if the things have the relevant properties (1990, 1996).

Bibliography

Works by Chisholm

1946: "The Contrary-to-Fact Conditional," *Mind* 55, pp. 289–307.
1948: "The Problem of Empiricism," *Journal of Philosophy* 45, 512–17.
1951: "Philosophers and Ordinary Language," *Philosophical Review* 60, pp. 317–28.
1952: "Comments on the 'Proposal Theory' of Philosophy," *Journal of Philosophy* 49, pp. 301–6.
1955a: "A Note on Carnap's Meaning Analysis," *Philosophical Studies* 4, pp. 87–9.
1955b: "Law Statements and Counterfactual Inference," *Analysis* 15, pp. 97–105.
1955–6: "Sentences about Believing," *Proceedings of the Aristotelian Society* 56, pp. 125–48.
1957a: "Chisholm–Sellars Correspondence on Intentionality," *Minnesota Studies in Philosophy of Science*, vol. II, pp. 511–20.
1957b: *Perceiving: A Philosophical Study*, Ithaca, NY: Cornell University Press.
1963: "Supererogation and Offence," *Ratio* 5, pp. 1–14.
1964: "J. L. Austin's Philosophical Papers," *Mind* 73, pp. 1–26.
1966: *Theory of Knowledge*, Englewood Cliffs, NJ: Prentice-Hall.
1974: "Practical Reason and the Logic of Requirement," in *Practical Reason*, ed. S. Körner, Oxford: Blackwell Publishers.
1976: *Person and Object*, La Salle, IL: Open Court.
1977: *Theory of Knowledge*, 2nd edn., Englewood Cliffs, NJ: Prentice-Hall.
1979a: "Objects and Persons: Revision and Replies," in *Essays on the Philosophy of Roderick M. Chisholm*, ed. E. Sosa, Amsterdam: Rodopi.

1979b: "The Indirect Reflexive," in *Intention and Intentionality*, ed. C. Diamond and J. Teichman, Ithaca, NY: Cornell University Press.
1981: *The First Person: An Essay on Reference and Intentionality*, Brighton: Harvester, and Minneapolis: University of Minnesota Press.
1982: *The Foundations of Knowing*, Minneapolis: University of Minnesota Press.
1986: "Self-Profile," in *Roderick M. Chisholm*, ed. R. J. Bogdan, Dordrecht: Reidel.
1989a: *Theory of Knowledge*, 3rd edn., Englewood Cliffs, NJ: Prentice-Hall.
1989b: *On Metaphysics*, Minneapolis: University of Minnesota Press.
1990: "Events Without Times: An Essay on Ontology," *Noûs* 24, pp. 413–28.
1993: "William James's Theory of Truth," *The Monist* 75, pp. 569–79.
1996: *A Realistic Theory of Categories*, Cambridge: Cambridge University Press.

Works by other authors

Gettier, E. (1963) "Is Justified True Belief Knowledge?," *Analysis* 23, pp. 121–3.
Lewis, D. (1979) "Attitudes *De Dicto* and *De Se*," *Philosophical Review* 88, pp. 513–43.
Quine, W. V. (1939) "A Logistical Approach to the Ontological Problem," *Journal of Unified Science* 9, pp. 84–9 (preprints only). (Reprinted in Quine 1976, pp. 196–202.)
——(1948) "On What There Is," *Review of Metaphysics* 2, pp. 21–38.
——(1960) *Word and Object*, Cambridge, MA: MIT Press.
——(1976) *The Ways of Paradox*, revised edn., Cambridge, MA: Harvard University Press.
Russell, B. (1910) "On the Nature of Truth and Falsehood," in *Philosophical Essays*, London: Longmans, Green & Co.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

23

Donald Davidson (1917–)

ERNEST LEPORE

Donald Davidson is one of the most important and influential philosophers of the second half of the twentieth century. He has never attempted a systematic exposition of his philosophical program, so there is no single place a student, interpreter, or critic can seek its official formulation. His published essays, taken together, form a mosaic that must be viewed all at once in order to discern an overall pattern. In addition, they sometimes exhibit an enigmatic quality, with subtleties, complexities, and cross-references that often cannot be entirely appreciated except in conjunction with each other. All of this can render access to his thought not just difficult but at times frustrating, despite its obvious importance.

In an effort to help a novice or even someone already stumped, the following major themes in Davidson's philosophy will be summarized: (1) Reasons and Causes, (2) Events and Causation, (3) Anomalous Monism, (4) Theory of Meaning and Compositionality, (5) Radical Interpretation, (6) Adverbial Modification, (7) The Method of Truth in Metaphysics, (8) Against Facts, (9) Truth and Correspondence, (10) Animal Thought, (11) Alternative Conceptual Schemes, (12) Anti-skepticism, (13) Anti-Cartesianism and First Person Authority, (14) The Rejection of Empiricism.

Though this list is not exhaustive, it identifies broad themes that structure Davidson's project. Of course, nothing short of reading Davidson's own work can supplant a detailed characterization of his take on each individual point, but I hope to tempt you to journey into Davidson's philosophical world.

Reasons and causes

We often try to explain another's actions by citing her reasons for so acting. Mary offers Bill the seat next to her on a train. She does so because she wants Bill to sit down and believes that by identifying this open seat he will. She thereupon asks him to sit down. What relation must obtain between Mary's behavior and her reason for this behavior in order to correctly conclude that she acted as she did *because* of her reason?

Until Davidson's "Actions, Reasons, and Causes" (see 1980), as difficult as it now seems to comprehend, something close to a consensus had formed in philosophy that whatever the relationship might be it could not be causal. It was believed that an alleged "logical connection" between reasons and actions excluded any causal relation

between them. The central purpose of Davidson's essay was "to defend the ancient – and commonsense – position that rationalization is a species of causal explanation" (1980: 3). Much of his essay is devoted to refuting various arguments, then popular, that purported to show that reasons could not cause the actions they rationalize. According to each, a necessary condition for causal interaction cannot be satisfied by reasons and actions. These arguments are too many to be properly treated here, but in passing it should be noted that they were inspired by remarks of the twentieth-century philosopher Ludwig Wittgenstein or by certain interpretations of the eighteenth-century British philosopher David Hume's strictures on causation.

By way of a single example, many authors believe that Hume had more or less established that in order for one thing *A* to be causally related to another thing *B*, *A* and *B* must not be logically connected. So, for example, when one billiard ball moving with a certain momentum hits another, then, unless circumstances are unusual, the second will be caused to move. The second ball's movement occurs not as a matter of logic, but rather as a matter of the physical nature of our universe. It is logically possible that causal interaction in this universe has been governed by a physical nature distinct from what actually governs it. One intuition behind denying that reasons can be causes of actions is that they are logically related; so, if I believe that smoking is harmful, and I desire no harm, mustn't I, as a matter of logic or of meaning alone, intend not to smoke? If so, then how could my reasons for intending not to smoke have caused me to have that intention?

For our purposes, it suffices to say that Davidson replied to this argument and others by showing either that reasons and actions indeed satisfy the necessary condition in question, or that the would-be necessary condition for causal interaction is in fact not one at all. So, with respect to Hume's observation, Davidson argued that a logical connection between descriptions of a cause and an effect does not by itself preempt causation, as is evident in "The cause of event *e* caused event *e*." No one would infer from the fact that an event could not have existed under the description "the cause of *e*" without being the cause of the event *e* that the first didn't cause the second. Similarly, even if no one could be described as having a belief under the description of "smoking is harmful" and having a desire under the description "don't do anything harmful" without having an intention under the description "don't smoke," it doesn't follow that there is no causal relation between these reasons and the intentional action that ensues.

Events and causation

In "The Individuation of Events" (in 1980), Davidson argues that we must recognize that we can describe the same action in different ways. Indeed, if we did not do this, we could not make sense of perfectly natural claims like "Jones managed to apologize by saying 'I apologize'," where a single action is described both as a managing to apologize and as a saying "I apologize." But what sort of object are these actions that admit of re-description?

Tables, chairs, and people are concrete, dated particulars, that is, unrepeatable entities with location in space and time. These features alone distinguish them from, say, either numbers or God. But what about actions and events, for example, the action

of Bob shooting Bill, or the event of Hurricane Floyd, or the event of the stock market crash of October 1929? Anyone who doubts these three events exist need only confer with one of their victims. But what sort of entity can an action or event be? We speak, for example, of the event of a baseball game between the Yankees and the Rangers as lasting three hours, and as taking place in New York. We might conclude that this event was exciting or too long or even crucial in determining a champion. On the basis of on such assessments what can we conclude about what sort of thing an event is?

Davidson's chief claim about actions and events is that like tables and chairs they are concrete, dated particulars that can be described in various nonlogically equivalent and non-synonymous ways. What distinguishes them from other sorts of concrete, dated particulars is their potential for causal interaction, and so, it's part of the nature of being an event that it can stand in a causal relationship. Davidson goes on to individuate them, so that two events are identical just in case they have the same causes and effects. Since Davidson treats causation as a relation between events, and takes action to be but a species of event, events comprise the very subject matter of action theory, as well as science and ethics. (We will take up below his argument for their existence and for specific claims as to their nature in the section on the method of truth in metaphysics.)

Davidson holds that events related by causation must be subsumable under some law or other, where a law is a generalization confirmable by its positive instances and, if true, supports counterfactual statements, and where an event is subsumed by a law just in case it instantiates that law (1980: 217). So, for example, according to Boyle's Law, the pressure of a fixed mass of gas at a constant temperature is inversely proportional to the volume of that gas. An instance of this law – if *a* is the pressure of a fixed mass of gas at a constant temperature, then *a* is inversely proportional to the volume of that gas – is *confirmed* just in case when its antecedent is true, so is its consequent. To say it *supports* its counterfactual instances means that even if *a* is not the pressure of a fixed mass of gas at a constant temperature, if it *were*, then it *would* be inversely proportional to the volume of that gas (1980: 215).

Davidson's views about the nature of events and their relation to laws brought him to a stunning conclusion about the relationship between minds and bodies, namely, his thesis of anomalous monism, to which we turn immediately.

Anomalous monism

Much can and has been said in favor of each of the following three claims:

1 The mental and the physical are distinct.
2 The mental and the physical causally interact.
3 The physical is causally closed.

The problem, though, is that they seem inconsistent. Consider their application to events. (1) says that no mental event is a physical event; (2) says that some mental events cause physical events, and vice versa; a loud noise reaching Tom's ear may cause him a desire to turn down his radio; and his desire to turn down his radio may cause his arm to move in such a way to result in the volume of his radio being lowered. (3) says that all the causes of physical events are themselves physical events. The dilemma

posed by the plausibility of each of these claims and by their apparent incompatibility is the traditional mind–body problem. Davidson's resolution, as articulated in "Mental Events" (in 1980), "The Material Mind" (in 1980), and "Philosophy as Psychology" (in 1980) consists of theses (4)–(6), which taken together comprise his thesis of *anomalous monism*:

4 There are no exceptionless psychological or psychophysical laws, and in fact all exceptionless laws can be expressed in a purely physical vocabulary (1980: 214–15, 231).
5 Mental events causally interact with physical events (1980: 208).
6 Event *c* causes event *e* only if an exceptionless causal law subsumes *c* and *e* (1980: 208).

The thesis is monistic, since it assumes there is but one kind of stuff in the world, physical stuff, but it is anomalous, since although its monism commits it to physical and mental stuff being the same stuff, it denies that that there is a strict reduction of the one to other. A full-blown exegesis and defense of this view and the claims which comprise it is beyond the intended scope of this essay, but a few words about the extraordinary thesis (4) are in order.

Thesis (4) is a version of (1). It is commonly held that whatever property a mentalistic predicate M expresses is *reducible* to one expressed by a physical predicate P (where "M" and "P" are not logically connected) only if an exceptionless law links them (where an exceptionless law is exactly what it says, namely, one that under no conditions admits of exceptions. Boyle's Law, stated above, is obviously not an exceptionless law). According to (4), mental and physical properties are distinct. In sketch, Davidson bases his argument against the possibility of exceptionless laws linking mental and physical predicates on their sufficiently distinct *constitutive* principles (1980: 222, 238).

Measurability of length, mass, temperature, and time are *constitutive* of the physical, inasmuch as these features govern the applicability of physical predicates (1980: 221). So, for example, anything physical must have length. Suppose, though, upon investigation, we discover that among three physical items, though the first is longer than the second and the second longer than the third, the first is *not* longer than the third. What would we conclude? We would assume that either we are mistaken, or that their lengths changed during the course of measurement. What we would not conclude is that the transitivity of length is false. Why? Because, no three things could have a physical predicate true of them unless their lengths respect transitivity; that is, respecting this constraint is constitutive of being physical.

With mental items, their constitutive principles include principles of rationality, for example, constraints about consistency and rational coherence (1980: 236–7). For example, the transitivity of desire – if a person desires *a* over *b* and *b* over *c*, he *ought* to desire *a* over *c*; or the consistency of belief – we assume that if an agent believes that *p*, he ought not also to believe that not *p*. This ties in with earlier discussion of reasons and actions; interpreting the behavior of another requires attributing beliefs and desires. These attributions are intended to provide an agent's rationale for acting, but they fail this task unless a degree of *rational* choice is presumed. By virtue of this presumption, the constitutive principles of the mental include *norms* of rationality. Davidson claims such constitutive principles have "no echo in physical theory" (1980:

231). So, he concludes, in an important sense psychology cannot "be reduced to the physical sciences" (1980: 259), namely, exceptionless laws cannot link the two sorts of sciences, because the normative relationships among mental states cannot be expressed in a physical language.

This compact discussion of *anomalousness* leaves a host of questions unanswered, but it serves to identify points of controversy surrounding Davidson's thesis (4). What about his thesis (6)? In "Causal Relations" (in 1980), Davidson argues that the most plausible interpretation of singular causal statements like "The short circuit caused the fire" treats them as two-place predicate statements with their singular terms, in this case, "the short circuit" and "the fire" designating events. Thesis (6) says that an event *c* causes an event *e* only if there are singular descriptions D of *c* and D′ of *e*, and an exceptionless causal law L such that L and "D occurred" entail "D caused D′" (1980: 158). But (6) and the second part of (4) entail that physical events have only physical causes, and that all event causation is physically grounded.

Given the parallels between (1)–(3) and (4)–(6), it may seem that the latter, too, are incompatible. Davidson, however, argues that they can all be true if (and only if) individual (token) mental events are identical to individual (token) physical ones (1980: 215, 223–4). Suppose an event *e* is physical just in case *e* satisfies a predicate of (basic) physical science, where such predicates are those that occur in exceptionless laws. Since Davidson assumes that only physical predicates (or predicates expressing properties reducible to physical properties) occur in exceptionless laws (1984b: 240), it follows that every event that enters into causal relations satisfies a physical predicate. But, then, it follows that a mental event that enters into a causal relation must satisfy some physical predicate, and so, is itself a physical event. His argument, if sound, establishes no more than that every concrete mental event is identical to some concrete physical one. Since this identity of mental and physical event tokens is compatible with rejecting systematic laws bridging mental and physical event-types, that is, with anomalous monism, the latter thesis only partially endorses (1). The mental and physical remain type-distinct insofar as mental and physical events are *not* linked by exceptionless laws under mental and physical descriptions.

Davidson's account of reasons and events impacts on, and is in turn, affected by, his radical approach to language; to see this, we turn now to examine his theory of meaning.

Theory of meaning and compositionality

Our discussion will be limited primarily to ideas present in "Theories of Meaning and Learnable Languages" and "Truth and Meaning" (both in 1984b), where Davidson identifies an adequacy criterion for theories of meaning for natural languages, and then applies it critically to a number of then prominent analyses of aspects of natural language (see TARSKI, CHURCH, GÖDEL). He also sketches a program in which, surprisingly, a certain austere style of meaning theory meets this adequacy criterion.

What aim(s) should a theory of meaning seek to accomplish? That will depend on which linguistic aspects a theorist wants to explain. So, for example, though natural languages are spoken by finite speakers without magical abilities, they still have an infinity of meaningful (non-synonymous) sentences, each of which, at least potentially,

a speaker could understand (at a given time). For any (indicative) sentence S of English, a new one can be formed by prefacing it with "It is believed that." For any two (indicative) sentences, S and S′, a new one can be formed by disjoining them with the word "or"; and so on for other productive mechanisms of our language. The novel sentences which these productive mechanisms give rise to are intelligible to normal speakers if their components are. This capacity seems to require that speakers have learned (a finite number of) rules that determine from a finite set of *semantic primitives* what counts as meaningful compositions (where an expression is semantically primitive if the "rules which give the meaning for the sentences in which it does not appear do not suffice to determine the meaning of the sentences in which it does appear" (1984b: 9)).

On the basis of such considerations, Davidson requires of a theory of meaning that it specify what every sentence means by exhibiting its meaning as a function of the meaning of its significant parts (based, presumably, on their arrangement in the sentence). Let's call any such theory for a language "a compositional meaning theory" for that language. Davidson was the first philosopher to bring to prominence the importance of the requirement that a theory of meaning of our language exhibit it as compositional (1984b: 23). The requirement focuses attention on the need to uncover structure in natural languages. While this is clearly something that philosophers from time immemorial have engaged in, this project had not been, until Davidson, clearly separated from the ancient but now discredited project of conceptual analysis.

Davidson's positive suggestion for a compositional meaning theory for a language L, surprisingly, utilizes no concept of *meaning* that goes beyond truth. To wit, his theory of meaning takes the form of a (finite) theory *of truth* that, for each sentence S of L, entails what we shall call a T-sentence of form

(T) S is true in L if, and only if, *p*,

where "*p*" specifies (in a metalanguage) conditions under which S is true in L. So, for example, an adequate compositional meaning theory for German should issue in a theorem like (S):

(S) "Schnee ist weiss" is true in German if, and only if, snow is white.

Why does Davidson choose this rather austere form of theory over, say, one that explicitly invokes meaning by issuing in theorems of the form

(M) S in L means that *p*,

where "*p*" specifies (in a metalanguage) what S means in L? His reasons have nothing, as sometimes suggested, to do with replacing the complex notion of meaning with one more tractable or easily understood. Davidson's inquiry is not guided by conceptual or metaphysical qualms about the notion of meaning, but solely by the goal of devising a compositional meaning theory. He argues that this aim can be achieved with a compositional meaning theory that issues in theorems that take the form of (T), but not by one that issues in theorems of form (M). (His reasons are based on the fact that unlike the locution "is true if and only if," the locution "means that" is semantically opaque, thus hindering the development of a compositional meaning theory. What semantic opacity is will be discussed below in the section on animal thought.)

A compositional theory of meaning for a language L that issues in interpretive T-sentences like (S) is such that anyone who knows it is positioned to understand every sentence of L. By specifying the meaning of a sentence S in L *via* sentences of form (T), Davidson is requiring a specification that enables anyone who understands the language in which the specification is given (the *metalanguage*) to understand (the object language sentence) S. The observation that natural languages are compositional is the foundation upon which Davidson builds his program in the theory of meaning.

Since we do not and could not know a priori how to interpret the expressions of a natural language or how to assign them interpretive truth conditions, an adequate compositional meaning theory must be *empirical*. Any such theory must function as a theory for a natural language. In the case of one's own language, of course, though what one knows about it is not a priori, no difficulty arises in identifying which sentences of form (T) are interpretive. The problem for a theorist concerning a language he already understands is simply to figure out what the axioms of the theory ought to be in order to construct proofs of them. A quite different problem confronts a theorist for foreign languages (even to some degree for other speakers of one's own language, a theme emphasized more in Davidson's later work). This is where the important notion of *radical interpretation* enters into his discussion.

Radical interpretation

What justifies a choice of one compositional meaning theory for a language over another? Davidson argues that an adequate compositional meaning theory must be empirically warranted under the practice of *radical interpretation*. What this means is that certain specific empirical considerations must be respected in choosing between distinct but true compositional meaning theories, namely, in opting for a compositional meaning theory for German that issues in (S) over one that issues in (W).

(W) "Schnee ist weiss" is true in German if, and only if, grass is green.

(W) is, as a matter of fact, true, but, unlike (S), it fails to *interpret* "Schnee ist weiss," and so no compositional meaning theory for German that issues in (W) can be adequate. But for languages we do not already understand, a compositional meaning theory must be selected on the basis of "evidence plausibly available to an interpreter," that is, "someone who does not already know how to interpret utterances the theory is designed to cover" (1984b: 128).

Davidson boldly claims that nothing can be a language unless a correct compositional meaning theory that issues in a true and interpretive sentence like (S) for each sentence of that language can be selected on the basis of the sorts of observations plausibly available to a radical interpreter. In a bit more detail, a radical interpreter, by definition, is ignorant of the language she is trying to interpret, and also lacks access to bilingual informants, prior dictionaries, and the like. A radical interpreter is generally allowed to be able to determine when an informant holds a sentence true, even though she fails to understand whatever sentence is being held true (1984b: 135). So, in effect, the primary data for radical interpretation are formulable by what might be called *singular held true sentences*, for example, (E):

(E) Kurt belongs to the German speech community and Kurt holds true "Es regnet" on Saturday at noon and it is raining near Kurt on Saturday at noon.

Data like (E) can be collected from a variety of speakers across a variety of times by someone who does not already understand German, and will eventually confirm a generally held true sentence like (GE):

(GE) For any speaker of the German speech community and for any time, that speaker holds true "Es regnet" at that time if and only if it's raining near him or her at that time.

Claims like (GE) provide evidence for a radical interpreter that speakers of the German speech community *take* some form of words to express a specific truth (1984b: 135). But what licenses inferring from data like (GE) a target theorem like (R)?

(R) "Es regnet" is true in German of a speaker *s* at a time *t* just in case it's raining near that speaker at that time.

Could all German speakers get it wrong? Davidson denies any such possibility, answering that the inference from (GE) to (R) is legitimate, because a certain *principle of charity* is legitimately presupposed (1984b: 137). According to this principle, the favored compositional meaning theory for a language L must entail sentences of form (T) such that most sentences speakers of L hold true are in fact true.

Under radical interpretation, sentences speakers hold true keep turning out to be true. This is no accident; rather, it's because radical interpretation is a special sort of project, namely, one that is constituted by this principle of charity. So, once data like (GE) are collected, we can infer its corresponding T-sentences (R) *via* this principle of charity. To boot, by virtue of securing a set of interpretive T-sentences for context sensitive sentences like "Es regnet," a truth theory automatically assigns interpretive T-sentences to context *in*sensitive sentences like "Schnee ist weiss." So, to a very rough approximation, suppose that, on the basis of what a radical interpreter collects, she devises for context sensitive sentences "Es schneit" and "Es ist weiss" interpretive T-sentences like (a) and (b):

(a) For any speaker and time, if the speaker utters "Es schneit" at that time, then the speaker's utterance of "Es schneit" is true just in case it's snowing near him or her at that time.
(b) For any speaker, time and object, if the speaker utters "Es ist weiss" at that time, then the speaker's utterance of "Es ist weiss" is true at that time of that object if, and only if that object is white at that time.

But on the basis of these generalizations, suppose that somehow or other a radical interpreter conjectures application conditions for sub-sentential expressions like "schnee" and "weiss," as in both (c) and (d):

(c) For any object and time, "schnee" is true of that object at that time if, and only if, that object is snow at that time.
(d) For any object and time, "weiss" is true of that object at that time if, and only if, that object is white at that time.

Such information we can imagine being exploited in attributing truth conditions (e) to a context *in*sensitive sentence like "Schnee est weiss."

(e) "Schnee ist weiss" is true if, and only if, snow is white.

The upshot is that only those compositional meaning theories that entail T-sentences licensed by radical interpretation are adequate.

In the next few sections, we will explore some philosophical ramifications of embracing truth theories as the correct form for compositional meaning theories, and of embracing radical interpretation as their presumed manner of confirmation.

Adverbial modification

We discussed Davidson's views about the nature and role of actions and events above. Yet the strongest argument he advances for the existence of events and certain views about their nature derives, remarkably, not from any piece of pure metaphysical reasoning, but rather from constraints on compositional meaning theories.

Any theory of meaning for a language must embody a view of the relationship between language and reality. Davidson's conviction is that a compositional meaning theory, by providing a view about this relationship, offers substantive answers to the metaphysical questions about reality and its nature. In particular, the best compositional meaning theory for English, for example, will require positing events in order to explain what we say for sentences about actions, events, and singular causal relationships (like "Frank's pushing Bill caused him to fall").

So, consider the action/event sentence (1) and an obvious candidate for its interpretive truth condition (2):

1 John hit Bill.
2 "John hit Bill" is true if, and only if, John hit Bill.

In (2), linguistic expressions are both used and mentioned. Both the words "hit," "John," and "Bill" and the corresponding aspects of the world, the people and the action which relates them, are discussed. In this limited sense, (2) could be said "hook up" language and reality. This hook-up remains silent, though, on the nature of reality, since it tells us no more than what the English sentence (1) requires for its truth, namely, that John hit Bill. However, since an adequate compositional meaning theory must be finite (1984b: 4–15) in constructing a compositional meaning theory for, say, English (an unbounded language), structure must be read into its sentences. Consider action sentences (3)–(5):

3 John hit Bill at six.
4 John hit Bill at six in the bedroom.
5 John hit Bill at six in the bedroom with the stick.

These are but three examples of how adverbial modifiers can be added, in the form of prepositional phrases, to an English sentence without compromising its grammaticality or intelligibility. There is no obvious specifiable upper limit upon the number of modifiers English allows us to sensibly attach to these sorts of sentences – "after dark," "on his ear," "on a Tuesday," and so on. Therefore, any compositional meaning that

treats each of such sentences as involving a distinct primitive relation threatens to offend against the finitude condition on a compositional meaning theory. Were we to try to devise a compositional meaning theory for English according to which (3) is true just in case the three-place relation of hitting obtains between John and Bill and six o'clock, what would we say about (4)? Is it true just in case the distinct four-place relation of hitting obtains among two people, a time and a place? And on it goes. This strategy for devising interpretive T-sentences, in effect, winds up treating each adverbial modifier as introducing a distinct and novel relation with a distinct number of *relata*. A compositional meaning theory prohibits positing indefinitely many distinct primitive predicates in a language. Recall, the aim of a compositional meaning theory is to explain how there can be indefinitely many non-synonymous meaning sentences given a finite basis of meaningful components.

The method of truth in metaphysics

On the basis of such considerations, Davidson advances a proposal which simultaneously reveals common elements in all these many distinct sentences, issues in their correct interpretive T-sentences, and validates logical implications among them, for example, that (4) logically implies (3), and that both (4) and (3) imply (1). His idea, roughly, is to interpret sentences like (1) and (3)–(5) in such a way that they are "revealed" to harbor an existential quantifier that ranges over events.

The metaphysical punch line is that events must exist, since, otherwise, a *finite* compositional meaning theory for English would be unattainable. (This is how Davidson justifies the sort of ontological commitment to events we saw at work in the section "Events and Causation.") More specifically, Davidson argues that the best compositional meaning theory for English will interpret action sentences like (1) and (3)–(4) roughly along the following lines of (1′) and (3′)–(4′):

1′ *There is an event* that is a hitting of Bill by John.
3′ *There is an event* that is a hitting of Bill by John *and* it occurs at six.
4′ *There is an event* that is a hitting of Bill by John *and* it occurs at six *and* it occurs in the bedroom.

In each interpretation, there is existential quantification over an event. Each invokes exactly the same three-place relation of hitting and thereby shows what we already intuitively recognize to be common ground among them. And in (3′) and (4′), adverbial modification is transformed into a predication of this posited event.

This strategy for discerning ontological commitments extends to any locution where quantification and predication are required in order to construct a satisfactory compositional meaning theory for natural language, and thus provides us with a general method for isolating ontological commitments for speakers. Note how far removed this style of argument is from a long-standing tradition of treating metaphysics as an independent discipline, which somehow abstracts from questions about meaning or about science, instead taking a middle ground between science and analysis for discerning the nature of reality. Davidson, perhaps, could be interpreted as challenging the need for any such middle ground.

Against facts

Adopting the method of truth in metaphysics as a strategy for discerning ontological commitment also has negative ramifications for traditional ontological commitment to facts. A view going back at least to early Russell and Wittgenstein of the *Tractatus* is that the world is populated with *facts* and that true sentences are true because they *correspond* to these facts. In more recent writings, for example, "The Myth of the Subjective" (1989b) and "The Structure and Content of Truth" (1990), Davidson attempts to refute the claim that sentences are representations of reality. In the latter essay, he argues against "the popular assumption that sentences, or their spoken tokens, or sentence-like entities or configurations in our brains can properly be called 'representations', since there is nothing for them to represent" (Davidson 1990). In particular, there are *no* facts. And if there are no facts, and so facts fail to make sentences true, then we can ask, In what sense are sentences representational?

On Davidson's approach to meaning, the best compositional meaning theory does not require treating sentences as corresponding to facts or propositions. (Look in (T) above, the candidate interpretation of "Schnee ist weiss." On its right hand side, there is an English sentence, which provides an interpretation of the German sentence without making any reference to a fact or proposition.) Though not all of what Davidson says in his attack on representations can be addressed here, enough can be said about his attack against facts and correspondence theories of truth to illuminate his methodological approach.

Davidson's main argument against facts and correspondence theories of truth appears in his "True to the Facts" (in 1984b). There we find his so-called Great Fact argument for the conclusion that, given certain plausible assumptions, there can be at most one fact. The assumptions Davidson presumes are "that a true sentence cannot be made to correspond to something quite different by the substitution of co-referring singular terms, or by the substitution of logically equivalent sentences." From these assumptions he argues that one can prove that "if true sentences correspond to anything, they all correspond to the same thing" (1990: 303). Rightly or wrongly, Davidson takes these assumptions to embody traditional wisdom about facts.

The main point of the Great Fact argument is that if a context satisfies these two assumptions, then that context is truth functional (that is, the context looks just to the truth-value of the elements it relates, rather than to richer features, such as their meaning). So, if one sentence is made true by some fact *f*, then every sentence that agrees in truth-value with it is also made true by *f*, and thus, the Great Fact.

Since Davidson places the concept of truth to the forefront of the project of giving a theory of meaning for a natural language, and since he rejects one traditional correspondence theory of truth, it's appropriate to ask what his view is about this central concept. In the next section, we will consider briefly Davidson's position.

Truth and correspondence

As noted in the last section, Davidson rejects the traditional correspondence theory of truth, correspondence with facts. Still, in a way, Davidson himself is a correspondence theorist, since he explains the truth of sentences in terms of a relation between lan-

guage and something else. Recasting a correspondence theory of truth this broadly, Davidson's claim is that the sort of compositional meaning theory for a natural language he endorses is a correspondence theory of truth (1984b: 70), because it explains what it is for sentences to be true, not by relating sentences to objects, but by relating predicates and referring terms to objects *via* the relations of satisfaction and reference, and exhibiting the conditions under which sentences are true in terms of those relations.

Recall from our earlier discussion of adverbial modification, that according to Davidson, the best compositional account for explaining the unboundedness of modification as in "John kissed Mary in the park after dark" will treat this sentence as *quantifying* over an event and predicating of this event that it is a kissing by John of Mary, and that it took place in the park and that it occurred after dark. If Davidson is right, then in order for this sentence to be true, there must be an entity in the world, an event, that, so to speak, makes this sentence true.

This way of thinking about correspondence promises to be more illuminating than the traditional correspondence theory, because we can exhibit by way of the proof of a T-sentence from the axioms of a truth theory how the truth conditions of the sentence are arrived at on the basis of the satisfaction conditions for its significant parts; for each non-synonymous sentence, there will be a different route to its truth conditions. Moreover, it is clear that this approach affords no way of eliminating the semantic concept(s) on which it is based, that is, satisfaction and reference, so this approach is not a redundancy theory. (The other side of this coin is that the approach sheds little light on what it is for a predicate to be satisfied by an object or for a singular term to make reference to an object.) It is in this sense, and only in this sense, that Davidson is a correspondence theorist. Davidson, in more recent work, rejects this label as misleading (1990), though not the view itself. It is worth noting that being a correspondence theorist in this sense is neutral with respect to traditional disputes between various brands of realism and anti-realism.

Thus far we have been exploring ramifications of Davidson's views about meaning and interpretation for the nature of an external reality; in the next few sections the discussion turns inwards to explore conclusions Davidson draws about an inner reality, about the nature of mind.

Animal thought

What comes first, language or thought? Intuition cuts both ways: it's hard to find a pet owner who believes that her pet is a thoughtless brute, regardless of whether that pet has any facility, to speak of, with language. If intuition has got this right, then language and thought are independent. Yet it seems almost equally intuitive, that at least when *we* think, we think in our native tongue. So at least for we who have language it seems as if the two are mutually dependent.

Davidson, however, defends the unrestricted thesis that *any* capacity to think requires facility with language, and so only creatures with a language can think. So much for pet lovers! In "Thought and Talk" (1984b) and "Rational Animals" (1986c), he begins his argument for his controversial thesis by noting that ascription of psychological states to others, e.g., belief, desire, intention, and the like, exhibit *semantic*

opacity. So, in attributing to a dog the belief that the cat went up the tree, would the propriety of this ascription be affected were we to substitute for "the tree" another expression that refers to the same tree, say, "the chestnut in your backyard"? Given that the tree *is* the chestnut in your backyard, would you be so inclined to ascribe to the dog that he believes that the cat went up the chestnut in your backyard? If not, this would disclose that your attribution of belief to the dog falls short of literalness; that is, it is semantically opaque (1986c: 475).

Davidson's main contention is that semantic opacity, a failure to preserve truth under co-referential substitution, exists only when language is tied to thought. He advances two different lines of argument for his conclusion. The first appeals to a *holistic* thesis about belief ascription, by which is meant that we could never have grounds for ascribing a single belief to an organism except against the background of a wide array of other beliefs. Since, as Davidson argues, we could never have grounds for ascribing the required array of background beliefs to creatures that did not have a language, we could never be warranted in ascribing to such creatures any thought at all. The argument runs as follows:

1 Belief ascription exhibits semantic opacity, and
2 semantic opacity requires that we regard beliefs as possessing some definite intentional content, and
3 the possession of a belief with a definite intentional content presupposes "endless" further beliefs (holism).
4 Therefore, a creature to whom we are warranted in ascribing a belief is one that must possess a sophisticated behavioral repertoire;
5 but only linguistic behavior exhibits the sort of complex pattern that might warrant such ascription.

This argument, even if sound, at most establishes that we are unlikely ever to have decisive *evidence* that a speechless creature has beliefs. But the views for which we can collect decisive evidences and what we can establish as truth can come apart. Davidson wants to draw a stronger conclusion, that "unless there is actually such a complex pattern of behavior, there is no thought" (1986c: 476). He does so by arguing that:

6 Propositional attitudes require a dense network of beliefs (holism), and
7 "in order to have a belief, it is necessary to have the concept of belief" and
8 "in order to have the concept of belief one must have language," that is, one must be a member of a "speech community." (1986c: 478)

The big question for Davidson is how to get from the ubiquity of belief (6) and the view that beliefs require second-order beliefs (7) to his conclusion that "a creature must be a member of a speech community if it is to have the concept of belief" (8) (1984b: 170). He argues as follows:

9 The possibility of belief or thought generally is taken to depend on the concept of a representation that might be true or false, and
10 a concept of truth and falsity includes some notion of an objective, public domain, and
11 this, in turn, is possible only for an interpreter. (in 1984b: 170, 1986c: 480)

Davidson holds that only utterances can afford the fine-grained structure required for attributing thought; for only a creature whose behavior exhibits the kind of structure implied by a compositional meaning theory is a creature in which semantically opaque representations can make an appearance.

Alternative conceptual schemes

Even if thought requires language, isn't it possible that different people, communities, cultures, or periods view, conceptualize, or make the world (or their worlds) in different ways? Couldn't another thinker have concepts or beliefs radically different from our own? Davidson, unsurprisingly, identifies conceptual schemes with sets of intertranslatable languages (1984b: 185). By so doing, he transforms the question about alternative conceptual schemes into one about whether there could be non-intertranslatable languages. But why should anyone believe that questions about conceptual relativity have anything to do with translation?

Davidson's identification requires two assumptions: first, that speakers alone have thoughts and second, that any concept a speaker possesses and any thought he can entertain is expressible in his language. (Both assumptions were evident in our discussion above of "Thought and Talk," and "Rational Animals.") Together these entail that a difference in the conceptual schemes of two people requires that a portion of the language that one speaks is not translatable into any portion of the other's.

Once conceptual schemes are identified with sets of intertranslatable languages, the question of whether sense can be made of radically different conceptual schemes reduces to whether sense can be made of two non-intertranslatable languages (or, much the same, to whether or not a "significant range of sentences in one language could be translated into the other" (1984b: 185)). Davidson argues that making sense of such talk requires a criterion for when a form of behavior can be counted both as speech behavior and as speech that is untranslatable into our own. He then argues that no sense can be made of a total failure of translatability between languages (1984b: 185), and so no one could be in a position to judge that others had concepts or beliefs radically different from his own (p. 197).

In short, pressures from the nature of radical interpretation together with the fact that "all understanding of the speech of another involves radical interpretation" (1984b: 125) force him to draw his critical conclusions.

Anti-skepticism

Another consequence of radical interpretation is anti-skepticism, that is, the impossibility of massive error. In a number of articles, beginning with "The Method of Truth in Metaphysics," and including "Empirical Content," and "A Coherence Theory of Truth and Knowledge," continuing more recently through "What is Present to the Mind," "The Conditions of Thought," "Three Varieties of Knowledge," and "Epistemology Externalized," Davidson argues, on the basis of his principle of charity, that an interpreter cannot find speakers to have largely false beliefs, even if she herself has no opinion as to the general truth and falsity of these beliefs. Given what beliefs are, and how their contents are determined on this story, Davidson is committed to the

impossibility that "all our beliefs about the world might be false" (1991b: 193). A radical interpreter must have beliefs about the world in order to succeed in ascribing to others beliefs about the world. But, as radical interpretation is conceived, she also must find others largely in agreement with her in those beliefs.

Davidson's anti-skeptical argument from radical interpretation rests on two assumptions, namely, that to be a speaker is to be interpretable by others, and that to be interpretable by others requires being largely right, not only in one's general beliefs, but in beliefs about the local environment. On the assumption that radical interpretation is possible, the proper way to state the requirement on a speaker is that her beliefs about her environment be mostly true. The crucial aspect of radical interpretation is the importance of *causality* in determining what someone means or believes. We cannot "in general fix what someone means independently of what he believes and independently of what caused the belief . . . The causality plays an indispensable role in determining the content of what we say and believe" (1986: 435). So, it is the central role of causation in fixing the contents of beliefs that ensures that the truth of everything we believe is not in general "logically independent" of having those beliefs; and that others cannot differ too much from us in what they believe.

The central claim throughout is that how the contents of beliefs are determined limits the extent of falsity and diversity discriminable in a coherent set of beliefs. In interpreting another, a radical interpreter ventures hypotheses as to what in the circumstances in question causes a speaker to hold true the sentence in question. This is supposed to provide him (*ceteris paribus*) with the meaning of that sentence. In every case there will be different causal chains leading to the same utterance. A radical interpreter must choose one. He does so by responding to something in the environment, and so converges on something that is a common cause both of his own response and of the utterance of the speaker, thereby correlating the two and thus giving the content of the speaker's utterances. (This is what Davidson calls "triangulation" (1986c: 480, 1991b: 159–60). The central thought that emerges is that without constraints on what a creature is thinking about in addition to those provided simply by treating it as a rational creature capable of thought and speech, any answer to the question what it is thinking about will be so wildly underdetermined that we can give no clear content to the idea that it is thinking anything at all. Is the dog thinking that the squirrel ran up the tree or instead that large brown thing in front of it or the tree that Uncle Bill planted ten years ago, and so on. Without any clear way to rule out one of these attributions over the others, why assume that the dog believes any one of them?)

The method of radical interpretation "enforces" on any successful interpreter the conclusion that a speaker's beliefs are largely true and largely like her own. Thus, global skepticism is ruled out.

Anti-Cartesianism and first person authority

A central feature of the Cartesian tradition in modern philosophy is that at the foundation of the structure of our justified beliefs about the world are our beliefs about our own mental states, our attitudes, experiences, and sensations. As we have seen, Davidson's approach both to meaning and interpretation, and to central issues in epistemology, is *anti-Cartesian* inasmuch as he rejects this assumption. A radical interpreter

is restricted to behavioral evidence in interpreting another. From this standpoint, Davidson treats the central concepts employed in interpreting another as theoretical concepts introduced to keep track of behavior. Viewed from his perspective, the role of a theory of interpretation is to identify and systematize patterns in the behavior of speakers in relation to their environment. If this is right, we do not first have access to facts about speakers' meanings and attitudes, including our own.

How he aims to reconcile his treatment of the central concepts of interpretation as theoretical with the presumption of first person authority, that is, with the fact that speakers are necessarily more authoritative in general about their own attitudes and sensations than others are, is a central topic in Davidson's later writings.

It seems scarcely intelligible that another could be as well placed as you are with respect to whether you believe that you are hungry, or in pain. This asymmetry in epistemic position is connected with a difference in the way we know our own mental states and the way others know them. In ascribing mental states to others, we rely on their behavior (or records of their behavior); but in the case of first person ascription, we do *not*. Indeed, in our own case we do not rely on evidence at all, and so do not consult our behavior. Although knowing something in a different way or not on the basis of evidence does not in itself guarantee that what is known is known better, we may expect that this difference in how one knows one's own mental states (first person knowledge) and how others do underlies first person authority.

The challenge that first person authority presents to Davidson's assumption of the theoretical character of the concepts of interpretation is first taken up in his "First Person Authority," where he offers an explanation for the presumption that "a speaker is right when he sincerely attributes a belief, desire, or intention to his present self" (1984a: 101) by grounding it in the assumptions that an interpreter must make in order to succeed at interpretation.

Davidson aims to explain the asymmetry between our knowledge of our mental states and our knowledge of the mental states of others (or, alternatively, the asymmetry between our own knowledge of our mental states and the knowledge others have of them) by explaining a closely related asymmetry: why there is a "presumption that a speaker is right when he sincerely attributes a belief, desire, or intention to his present self, while there is no such presumption when others make similar attributions to him" (1984a: 101). His explanation of first person authority rests on an explanation of an asymmetry between the knowledge a speaker and interpreter have of meanings of the speaker's words. This asymmetry between the knowledge one has of one's own words and an interpreter's knowledge of the meanings of one's words is most striking in the case of an interpreter who is not a member of one's speech community.

The form of Davidson's argument is the following: one speaks a language only if one is interpretable; one is interpretable only if one is mostly right about the meanings of one's words; therefore, one speaks a language only if one is mostly right about the meanings of one's words.

Davidson's central methodological assumption is that a third person point of view on others' utterances and psychological states is primary in the sense that behavioral evidence forms our only evidence for the application of linguistic and psychological concepts and terms to others, and that their content is to be understood wholly in terms

of their role in accounting for the behavioral evidence available to us from this standpoint. His shift of viewpoint is so fundamental that, once adopted, the whole landscape in the philosophy of language and mind looks different. If Davidson is right, the central mistake of our philosophical tradition is the assumption of the Cartesian standpoint, and, in particular, the central place our tradition accords to the epistemic priority of knowledge of our mental states to knowledge of the world and other minds. Once this assumption is relinquished, each domain in which we have knowledge will be seen as necessary for the others, but knowledge of the world and by extension of other minds will turn out to be autonomous from knowledge of our own minds, in the sense that it is not explicable by appeal to inferences from a basis in knowledge of our own minds. In light of the alternative, Davidson's picture is attractive. Part of its interest and power lies in its promise to lay to rest what have been perhaps the central problems of the tradition from the beginning of the modern period. Despite the difficulties it faces, it is worth pursuing.

The rejection of empiricism

Another consequence of his taking what we might call a third person perspective of the radical interpreter as methodologically fundamental is the rejection of all forms of traditional empiricism. Essential to traditional empiricism is its attempt to account for our knowledge of the world exclusively by appeal to sensory experience. What is distinctive about empiricism is not the thought that sensory experience can play a role in justifying our beliefs about the world around us, but that it plays the role of a foundation for our empirical knowledge. This in turn entails that the first person point of view is fundamental, since each person's experience is treated as being his own foundation for his empirical knowledge. In adopting the third person point of view as fundamental, then, Davidson rejects a central tenet of all forms of empiricism, and the traditional project associated with it of explaining our empirical knowledge by appeal to experience. Rather, in Davidson's view, our knowledge of the world around us, of other minds, and of our own minds, has a unified source in our nature as rational beings capable of communicating with one another.

In conclusion, Davidson argues that language, mind, and action are inseparable. To account for language, that is, to answer the question, What is meaning?, he advances the radical idea that a theory of meaning can be satisfactory only if it discovers a finite basic vocabulary and rules of composition in the language to be interpreted. The aim to provide a comprehensive understanding of natural languages led him to a treatment of the theory of truth for a language as an empirical theory, and to the adoption of the stance of the radical interpreter as the standpoint for confirmation, linking the structure of a rich theory with its basic evidence, and placing the theory of meaning in the context of a theory of rational agency. Adopting this stance as fundamental is tantamount to the rejection of Cartesianism and empiricism, and so the abandonment, among other philosophical mainstays, of conceptual relativism, global skepticism, and representationalism. Theories frequently yield insight into problems that they were not specifically designed to solve. As with other significant philosophers, a careful reading of Davidson's writings bears out both how broad in scope his philosophical accomplishments are and, more importantly, how well they cohere.

Note

Some of this entry, in particular, from the section "Against Facts" to the end, is adapted from Kirk Ludwig's and my forthcoming *Donald Davidson: Truth, Meaning, and Reality*, Oxford University Press.

Bibliography

Works by Davidson

1980: *Essays on Actions and Events*, Oxford: Oxford University Press.
The collection contains the following essays, listed here with their original publication dates, and page extents in the collection.
"Actions, Reasons, and Causes" [1963], pp. 3–19.
"Causal Relations" [1967], pp. 149–62.
"Mental Events" [1970], pp. 207–27.
"Psychology as Philosophy" [1973], pp. 229–44.
"The Individuation of Events" [1969], pp. 163–80.
"The Logical Form of Action Sentences" [1967], pp. 105–48.
"The Material Mind" [1973], pp. 245–59.
1984a: "First Person Authority," *Dialectica* 38, pp. 101–11.
1984b: *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press.
The collection contains the following essays, listed here with their original publication dates and page extents in the collection.
"In Defence of Convention T" [1970], pp. 65–76.
"On the Very Idea of a Conceptual Scheme" [1974], pp. 183–98.
"Radical Interpretation" [1973], pp. 125–39.
"The Method of Truth in Metaphysics" [1977], pp. 199–214.
"Theories of Meaning and Learnable Languages" [1965], pp. 3–15.
"Thought and Talk" [1975], pp. 155–70.
"True to the Facts" [1969], pp. 37–54.
"Truth and Meaning" [1967], pp. 17–36.
1986a: "A Coherence Theory of Truth and Knowledge" [1983], in LePore 1986, pp. 307–19.
1986b: "Empirical Content" [1982], in LePore and McLaughlin 1986, pp. 320–32.
1986c: "Rational Animals" [1982], in LePore and McLaughlin 1986, pp. 473–80.
1989a: "Conditions on Thought," in *The Mind of Donald Davidson*, issue 36 of *Grazer Philosophische Studien* (ed. Brandl and Gombocz), Amsterdam: Rodopi, pp. 193–200.
1989b: "The Myth of the Subjective," in *Relativism: Interpretation and Confrontation*, ed. M. Karusz, South Bend, IN: University of Notre Dame Press.
1989c: "What is Present to the Mind," in *The Mind of Donald Davidson*, issue 36 of *Grazer Philosophische Studien* (ed. Brandl and Gombocz), Amsterdam: Rodopi, pp. 3–18.
1990: "The Structure and Content of Truth," *Journal of Philosophy* 87, pp. 279–328.
1991a: "Epistemology Externalized," *Dialectica* 45, pp. 191–202.
1991b: "Three Varieties of Knowledge," in *A. J. Ayer: Memorial Essays*, ed. A. Phillips, Cambridge: Cambridge University Press, pp. 153–66.

Works by other authors

Fodor, J. and LePore, E. (1992) *Holism: A Shopper's Guide*, Oxford: Blackwell Publishers.
Hahn, L. (ed.) (2000) *The Philosophy of Donald Davidson*, La Salle, IL: Open Court.

LePore, E. (ed.) (1986) *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell Publishers.
LePore, E. and McLaughlin, B. (eds.) (1986) *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell Publishers.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

24

G. E. M. Anscombe (1919–2001)

ANSELM MÜLLER

Gertrude Elizabeth Margaret Anscombe, British philosopher, studied Greats at Oxford (1937–41), and went as a research student to Cambridge, where she became a pupil and close friend of Ludwig Wittgenstein. She was appointed Research Fellow (1946), Lecturer (1951), and Tutorial Fellow (1964) of Somerville College, Oxford. In 1967 she was elected Fellow of the British Academy. She held the Chair of Philosophy in the University of Cambridge from 1970 to 1986.

Her philosophical outlook has been influenced most of all by Aristotle and by Wittgenstein. She is one of Wittgenstein's literary executors, and has translated and edited large parts of his work. At the same time she shows great originality, not least in the way in which she brings Wittgenstein's ideas to bear on topics that he did not himself explore. Many of her papers are remarkable also for the uniquely appreciative, unsparing, and creative manner in which she engages with great minds of the past, such as Hume. Anscombe has a gift for spotting what is most basic in traditional problems, and often her solutions seem to open one's eyes to what lay under one's nose. Her language is forceful and austere, her thinking unrestricted by convention or fashion. An early example of her independence of mind can be seen in "The Justice of the Present War Examined," a pamphlet written with Norman Daniel in the autumn of 1939. Here she queried both the aims of the British Government, and the means likely to be deployed, in fighting the war against Germany; she already foresaw "area bombing," foresight that lay behind her opposition, in 1956, to the conferment on President Truman of an honorary degree by the University of Oxford (1981a, III: 72–81; cf. vii and 62–71).

Anscombe has contributed to all principal areas of philosophy. The following summary account of her work considers it under seven main headings, covering most of her published writings.

Language, thought, and reality

Apart from two articles entitled "Subjunctive Conditionals," which argue that "if–then" is, roughly speaking, truth-functional (in 1981a, II: 196–207), and "On Private Ostensive Definition," which expounds and defends the view that there can be no private conferment of meaning, no single book or essay of Anscombe's is a systematic

treatment of questions in the philosophy of language and logic. She contributes to it rather by showing up weak spots in received answers, taking as her clue, for the most part, passages from ancient Greek philosophers or Wittgenstein. Typical is her remark: "We are accustomed to think that Plato in the Cratylus was extraordinarily blind in assuming that phonemes have meaning-roles. But this, as often, may be a failure on our part to see a problem" (1981b: 150).

Thus, "Understanding Proofs," an imaginary continuation of *Meno* 85d9–86c2 (1981a, I: 34–43) can be read as a challenge to give a better account than the Platonic theory of reminiscence, of the fact that simple conceptual truths cannot be understood without being believed. Chapter 1 of *An Introduction to Wittgenstein's Tractatus* (1959) contains an elaborate argument, not to be found in the *Tractatus* itself, in support of the view that "a (very large) class of mutually independent propositions" is presupposed in the common explanation of truth-functional tautologies. In other parts of this book and in papers such as "Parmenides, Mystery and Contradiction" (1981a, I: 3–8) and "The Early Theory of Forms" (pp. 9–20) we find keen observations on issues such as negation and the internal structure of sentences, truth and falsehood, proposition and assertion, meaning and reference, use and mention, tense and modality, universals, classes, and predication.

Again and again, Anscombe returns to classical problems of how words and propositions, concepts and thoughts are related to the realities they signify. In an early masterpiece, "The Reality of the Past" (1981a, II: 103–19), she shows that the question "How is it that statements about the past have meaning*?*" must be answered by *describing the use of the past tense* rather than appealing to "the experience of remembering" or gesturing at the past thing "reached by thought" or "seen through" the present evidence. She is intrigued by the more general Parmenideo–Platonic problem: "How can we think what is not?", which she contrasts with the modern question: "How could there be what we cannot think*?*" The Parmenides paper points out that, while of course we cannot without absurdity *say* of anything that it is but cannot be thought, we may reasonably *suppose* that something is but cannot be thought. Concerning modern attempts "to deduce what could be from what could hold of thought," Anscombe believes that "the ancients had the better approach, arguing only that a thought was impossible because the thing was impossible" (1981a, I: xi).

Is Wittgenstein's later philosophy a version of the modern approach ("*essence* is expressed by grammar," that is, by the rules which govern our application of words to that whose essence is in question)? "No," we are told in "The Question of Linguistic Idealism" (1981a, I: 112–33), an exposition and defense of Wittgenstein's views on the relationship between language and reality (see WITTGENSTEIN). "It looks as if *either* the grammar corresponded to something of the object, its real essence, which it has whether there is language about it or not, *or* the 'object' were itself dependent on language" (p. 113); both seem unacceptable. Anscombe's solution is this: On the one hand, reality does not force on us the concepts in which we relate to it ("How could an experience dictate the grammar of a word?": p. 114). Sameness of experience, or of kind of object, cannot determine the shape of a concept, since it is the correct (re-)application of the corresponding expression which, in the first instance, settles which sameness – and hence, which experience or kind of object – we have in mind. (There may be a tension between this claim and the rhetorical question she quotes from Wittgenstein:

"Do we make a concept wherever we see a similarity?") Thus alternatives to our set of concepts are indeed possible. On the other hand, the existence of whatever our concepts apply to does not therefore depend on our being there to conceive of it. "These essences, then, which are expressed by grammar, are not created by grammar" (p. 114). But there is room for "a partial idealism" (p. 118), of which more in the section "Existence by Convention and Intention," below.

The second part of the paper adresses a further problem. The grammar of our language governs the permissibility of the judgments expressed in it, by laying down, in particular, what counts as decisive evidence for them. However, it lays down also what is taken for granted not on the basis of evidence but, for example, as implied in the ways we judge and argue and act, or as a result of teaching. The corresponding "hinge" propositions vary vastly in subject matter and role. Compare "My name is L. W."; "The earth has existed for a long time before I was born"; "There is no God"; and "Caesar is a historical figure" ("Hume and Julius Caesar": 1981a, I: 86–92). Can "assumptions" that are in this way at the bottom of a linguistic practice themselves be right or wrong? "Finding grounds, testing, proving, reasoning, confirming, verifying are all processes that go on *within*, say, one or another living linguistic practice which we have" (p. 130). And divergence in judgment on account of divergence in "world-picture" or "knowledge system" is not a matter of *mistake* but rather "disagreement in the language" used (p. 131). However, Anscombe also takes Wittgenstein to hold that someone who comes to jettison a certain kind of groundless assumption or its certainty may be *right or wrong* in believing he realizes that formerly he was not competent to judge. And from this she concludes: "*That one knows something is not guaranteed by the language-game*"; that is, even where the rules of our linguistic practice leave no room for doubt, falsehood is not excluded (p. 132f.).

Some of Anscombe's essays concern the notion of a material substance, defending its coherence, and criticizing "bare particular" conceptions, comparable misunderstandings of the notion of matter, and empiricist objections and alternatives to a basically Aristotelian view. (See "The Principle of Individuation," 1981a, I: 57–65; "Substance," II: 37–43; and "Aristotle: The Search for Substance," in *Three Philosophers,* 1961 (together with P. T. Geach), pp. 3–63.

Time, necessity, and causation

Anscombe's "first strenuous interest in philosophy was in the topic of causality," and much of her later work in this area is an elaboration of the idea that the future is undetermined in the sense that, for example, there is "no such thing as how someone would have spent his life if he had not died a child" (1981a, II: vii). "Aristotle and the Sea Battle" (I: 44–56) already shows her an incompatibilist: "If what the typewriter is going to do is necessary, I cannot do anything else with the typewriter" (p. 48). Anscombe maintains Aristotle's view, canvassed earlier in "The Reality of the Past" (1981a, II: 112–16), "that nothing whatever could make what is certain untrue" (I: 52); and, for this reason, that "when *p* describes a present or past situation, then either *p* is necessarily true, or *-p* is necessarily true" (p. 53).

This kind of necessity is further explored in the celebrated 1971 inaugural lecture at Cambridge University, "Causality and Determination" (1981a, II: 133–47). From

Aristotle onwards, almost all philosophers (including Hume) have seen the essence of causality in necessitation. On Anscombe's alternative account, the notion of cause is embodied, in the first instance, in the use of such verbs as "*scrape*, *push*, *wet*, *carry*, . . ." (p. 137). This notion is one of *A deriving, or coming, from B*: of something that (*pace* Hume) is often observable. And "if *A* comes from *B* this does not imply that every *A*-like thing comes from some *B*-like thing or set-up or that every *B*-like thing or set-up has an *A*-like thing coming from it; or that given *B*, *A* had to come from it, or that given *A*, there had to be *B* for it to come from. Any of these may be true, but if any is, that will be an additional fact, not comprised in *A*'s coming from *B*" (p. 136).

The second part of the lecture examines the notion of determination and its applicability to physical events. "When we call a result determined we are implicitly relating it to an antecedent range of possibilities and saying that all but one of these is disallowed . . . [by] . . . something antecedent to the result" (1981a, II: 141). We may know that A has been *caused* by B without having any reason to think it was, in that sense, *determined* by B or anything else. A *system* like Newtonian mechanics would, it is true, provide such a reason; and the solar system offers a misleadingly undisturbed instantiation of its laws, which can make it look as if all causality had to match this model. But this appearance is illusory on two counts.

(1) Only if this system (or a comparable one) applied to arbitrarily small quantities (so as to exclude even minute causal indeterminacies and their multiplication over time) – only then would the result of, say, many balls interacting with each other for some time in stable surroundings, be (not only caused but) determined by some initial situation. And where we cannot assume such a system, we shall have to admit non-necessitating causes, like the radioactive material which (in Feynman's thought experiment) *may or may not*, via some Geiger counter, *cause* a bomb to go off (pp. 144f.). Anscombe hopes that this particular kind of example may prevent us from going on "as if undeterminedness were always encapsulated in systems whose internal workings could be described only by statistical laws, but where the total upshot, and in particular the outward effect, was as near as makes no difference always the same" (pp. 146f.).

(2) Even a physicist who believes that "the result that happens ought to be understood as the only one that [in the circumstances] was possible before it happened" (1981a, II: 142) need not be a *determinist*. Suppose Newton's laws were valid for arbitrary quantitative dimensions and thus provided for necessitating causes. This supposition does not yet rule out prevention and interference from other forces, and, in this sense, the possibility of alternative results. Determinism involves more, namely the belief that "the whole universe is a system such that, if its total states at t and t' are thus and so, the laws of nature are such as then to allow only one possibility for its total state at any other time." Anscombe sees no reason for believing this and, moreover, thinks it incompatible with the freedom of action, which after all for the most part concerns physical movements: "if these . . . are physically predetermined by processes which I do not control, then my freedom is perfectly illusory" (p. 146).[1]

In other works, Hume's account of causality is found wanting on the following three counts. (1) He argues that it is imaginable and therefore possible for a beginning of existence not to have a cause. But, Anscombe asks, can we determine, without

identifying a cause, that there and then an object *started to exist*, rather than *arrived* (having travelled "say as a gas": 1981a, II: 161)? (2) Hume fails to tell us what kind of specification of a cause and its effect is to count if we want to convince ourselves of his claim that the idea of the one is "distinct" from that of the other (p. 150). (Such distinctness *seems* to be absent, for example, when X's *mother* is said to be the cause of X.) (3) I may know without observation of something that it has caused me to do something, as when "I thought I saw a face at the window and it made me jump" (p. 75). This type of cause (which Anscombe calls a "mental cause") does not lend itself to Hume's explanation in terms of regular succession at all.

From experience to self-consciousness

Anscombe is best known for her influential work in the philosophy of mind. Apart from her classic *Intention*, she has produced "case studies" such as "The Subjectivity of Sensation" (1981a, II: 44–56), "Comments on Professor R. L. Gregory's Paper on Perception" (pp. 64–70), "On Sensations of Position" (pp. 71–4), "Pretending" (pp. 83–93, including an account of *anger*), "On the Grammar of 'Enjoy' " (pp. 94–100), and "Will and Emotion" (1981a, I: 100–7). One of her main targets is a temptation to bring everything mental under the heading of "experience," thus assimilating it to sensations and images. (See "Events in the Mind": 1981a, II: 57–63.) In fact, a psychological concept may make essential reference to a wide variety of things such as antecedent and surrounding conditions, behavioral and verbal expression, actions and aims, assumptions and thoughts, capacities and tendencies.

In "Memory, 'Experience' and Causation" (1981a, II: 120–30), after reminding us that a mental image cannot be the essence of memory because (memory) *beliefs* "are involved in referring an image to the past" (p. 126), Anscombe argues that there is no *core experience* to memory *at all* (as there is to seeing, or hearing). Genuine remembering is generally assumed to be composed of some such experience, M, plus an "appropriate" causal link between what is remembered and M. Consider, however, the case of X, who knows about a past event, E, in his life without knowing whether he does so because he remembers E or because he has been told of it. Interestingly, *this* lack of knowledge does not preclude X from *knowing* that E happened. Nor, Anscombe argues, does it consist in X's failing to know whether his present certainty of E was *caused* by E. Suppose now that X's knowledge is *not* in fact *memory* of E; then, of course, M does not come into the picture at all. If, on the other hand, it *is* a case of memory ("he must be remembering"), again it cannot involve M; for M was supposed to be a "memory experience," leaving no room for questions like "Is my belief, if true, a case of remembering E?" Hence M is a philosophers' fiction. As Anscombe points out, this is a problem not only for Cartesians and Empiricists, but also for any materialist identity theory of the mind (p. 128).

According to Anscombe, the human mind is characterized not so much by experience as by self-consciousness, understood not as some kind of self-perception but as the basis of self-ascription of all sorts of things *including* experiences. "Self-" here has the same function as in the claim that everyone uses "I" to speak of him-, or herself and, indeed, as the word "I." Does this stand for the human being who uses it, or rather for a Cartesian ego? Neither, we are told in "The First Person" (1981a, II: 21–36). "I" is

not, since it has not the sense of, a name, demonstrative, or other "referring expression." Such a sense would require "a 'conception' through which it attaches to its object" (pp. 28f; cf. n. 2). For the same reason, NN, in saying "I am NN," may express knowledge (or a mistake, or lie) but does not make a statement of identity. Suppose that everybody could avail themselves of every possible conception of every human being. The resulting information would leave unanswered for me but not for others the question *which one of these people I am*.

Human action and practical thought

Intention first appeared in 1957 and has been influential ever since, perhaps even more than is generally recognized. We speak of intention to ϕ, intentional ϕ-ing, and (further) intention in ϕ-ing. The second of these forms is the central topic of Anscombe's inquiry. Intentional actions can be marked off from non-intentional actions as those which the agent, without relying on observation, knows he is performing and to which he allows the question "Why?" to apply in a special sense. This sense is marked off from the sense assumed in an answer that would state a cause and, in particular, a mental cause. A relevant answer will give a *reason* for ϕ-ing which either looks backward (as in revenge: "I ϕ because he hit me") or amounts to a further intention (an interpretative reason: "My ϕ-ing is a signal for NN"; or an end: "I am ϕ-ing in order to find X"); or it will indicate that there is no reason ("I just thought I would ϕ"). The latter case must be distinguished from a *rejection* of the "Why?" question ("I was not aware I was ϕ-ing" or "I observed that I was ϕ-ing" or "I don't know the cause").

Knowledge of what you are doing, not by observation but *in intention*, may be called "practical knowledge." Discrepancy between thought and reality is here blamed on the latter. If you do not buy what is on your shopping list, the mistake is in your performance not in the list, as it *would* be in the list of a detective who made a mistake in tracking your proceedings. What you practically know you are doing will often coincide with the conclusion of an Aristotelian "practical syllogism." If this conclusion is *to* ϕ (or words such as "So I'll ϕ"), your premises mention (1) something wanted (under some "desirability characterization" or other) and (2) ϕ as a way of achieving it. Such practical reasoning need not "necessitate": ϕ-ing may not be the *only* way of realizing the thing you want. Also, the connections exhibited in a practical syllogism may truthfully be stated in your answer to the "Why?" question even though you did not go through them before you ϕ-ed. These ideas are further developed in "Thought and Action in Aristotle: What is 'Practical Truth'?" (1981a, I: 66–77), "Practical Inference" (1995: 1–34), and "Von Wright on Practical Inference" (1989: 377–404).

Under a description

As Anscombe points out (1963: 37–49), one and the same action may be intentional under some descriptions and not under others. In a single action you may be intentionally moving your arm *and* intentionally cutting bread but unintentionally (though *perhaps* knowingly and therefore voluntarily) contracting these particular muscles and pointing towards X with the bread-knife. In "The Two Kinds of Error in Action" (1981a,

III: 3–9), this is brought to bear on the legal and moral assessment of actions. Since, for example, "you may consent to something under one description and not under another, the fact of fraud may be a proof that a *certain* consent has not taken place at all" (p. 3). And "if a man genuinely and reasonably, but wrongly, thought that this was property he had a right to take away, then we say 'That was not stealing at all' " (p. 5). More generally, culpable and non-culpable ignorance, knowledge and intention can be relevant in various ways to (1) what descriptions are true of an action of yours, and (2) whether you are responsible for the action under a given description which applies to it. Here also belongs the distinction, essential to the doctrine of "double effect" (see below), between descriptions under which you intend an action and descriptions under which you merely know it to be involved in what you are doing intentionally.

In "Under a Description" (1981a, II: 208–19), Anscombe defends this expression against attacks and misunderstandings. She extends its application and reminds us that "the description *under which* [something] is aimed at is that under which it is *called* the object" (1963: 66). The description under which an action is intended/something is an intentional[2] object (e.g. of sight), can be (1) non-interchangeable, (2) indeterminable, and (3) existentially non-committal, in ways that may be elucidated by the following examples. (1) "I meant to cut bread but not to point the knife towards X."/"Didn't you see the blood?" "Well, I saw red patches on the floor." (2) Cutting a slice roughly 1 cm thick will *be* cutting it 0.90 or cutting it 0.91 or . . . cm thick. But intending to cut it roughly 1 cm thick is not intending to cut it 0.90 or intending to cut it 0.91 or . . . cm thick./The people you saw were thirteen in number; but as an eye-witness you may *have to* stick to "I saw quite a lot of people" or "perhaps a dozen." (3) "Cutting bread" describes my intention but not what I am actually doing if the knife is hopelessly blunt./"Was there a real flash of light when *I saw one*, or was there something wrong with my eyes?" (cf. p. 4).

In "The Intentionality of Sensation" (1981a, II: 3–20), Anscombe shows how different accounts of perception suffer from a common neglect of this topic: phenomenalism "misconstrues intentional objects as material objects of sensation," while "'ordinary language' philosophy . . . does not allow for a description of what is seen which is e.g. neutral as between its being a real spot (a stain) or an after-image" (pp. 11f.). In "Causality and Extensionality" (pp. 173–9) she comments on the inadequacy of a causal statement like "The child died because the tallest girl in town is Rhesus-negative" as compared with ". . . because his mother is Rhesus-negative." On Anscombe's view, such statements should be understood as non-extensional statements connecting, non-truth-functionally, genuine propositional components. Here again, we may say that it is the *description under which* the elements of cause and effect are identified that matters to the intelligibility if not truth of the causal claim.

Existence by convention and intention

Alternative descriptions are in play also where, in virtue of given conventions, a certain distribution of ink on a piece of paper is an English sentence, or a particular killing capital punishment. "Convention" here points to cultural constitution rather than agreement, let alone arbitrariness. It is a pervasive theme in Anscombe, its variations surfacing in the philosophy of (1) language, (2) knowledge, (3) action, (4) morality, and

(5) social institutions. The extensive relevance of the topic is made explicit in "Rules, Rights and Promises" (1981a, III: 97–103), where Anscombe mentions the "natural unintelligibility" (Hume's phrase) of promises, contracts, rights, legal obligation, etiquette, rules of games, rules of grammar and logic, infringement, and sacrilege.

(1) In "A Theory of Language" the question "*what* about the occurrence of a sound constitutes it a sign" (1981b: 150) is left unanswered. But language-game descriptions are said to give us, by way of comparison and without recourse to the notion of meaning, an idea of the possible functioning of a word in use: an idea of how the grammatical conventions of our actual language work. There is no right or wrong about these conventions. They create our concepts, but not what these are concepts *of* (cf. the section "Language, Thought, and Reality," above), with the notable exception of "promises . . . rules and rights, [which] are essences *created* and not merely captured or expressed by the grammar of our languages" (1981a, III: 100; see also (3) below).

(2) Grammar is, however, supposed to determine not only criteria (the type of evidence for the presence of X whose prima-facie validity is part of the concept of X), but also standards of comparative certainty for potentially incompatible judgments. Is convention at the bottom of these standards, too? And is there, in case of conflicting standards, any court of appeal? These questions (also discussed in the first section, above) do not seem to receive a definitive answer in Anscombe's work.

(3) In a short paper "On Brute Facts" (1981a, III: 22–5) we are introduced to the idea of facts, describable as A, which (in a society with certain institutions, given a vaguely specifiable normal context and the absence of an indefinite range of defeating conditions) "amount to" facts describable as B. For instance, making the above assumptions, the fact (A) that X has delivered a quarter of potatoes to Y amounts to (and is "brute" relative to) the fact (B) that Y owes money to X. But *how* can, in this case, the event described as A constitute the obligation claimed in B? How can, in a whole area of comparable cases, an "ought" derive from an "is"?

These questions are taken up in an essay "On Promising" (1981a, III: 10–21). Under suitable conditions, my uttering "I promise you to ϕ" amounts to a promise to ϕ. One of these conditions is my intention to promise. But (a) how can we invoke this intention in explaining *promising to* ϕ, if an account of the intention has to mention its content, i.e. that very promise? And (b) how can I, merely by uttering certain words with that intention, bring about restrictions, which did not exist before, on my possibilities of acting? Anscombe's answer to both questions consists in a highly original application of Wittgenstein's idea of a language-game. She imagines the following *practice* (pp. 15–17): There is a form of words "Bump! I'll ϕ." A participant NN who has used it is liable to be made by others to ϕ. The pressure they put on NN may be physical or, at a less primitive stage of the language-game, conventional. In the latter case, they address "stopping modals" to NN, like "you can't" and "you have to." These "are at first words used by one who is making you do something (or preventing you), and they quickly become themselves instruments of getting and preventing action" (p. 101). Finally, they are combined with "logoi" like "(but) you bumped to ϕ!" (pp. 101f., 142f.): "reasons," whose connection with ϕ-ing "*is itself nothing, except that it is linguistically MADE*" (p. 140). If recalcitrant, NN is reproached for *having used those words*

and not ϕ*-ed*. This practice of bumping has the significance of promising, if we assume that a participant will *try to extract* a "Bump! I'll ϕ" from others when he *wants* them to ϕ, and use their having bumped to ϕ as a weapon in making them ϕ, etc.

How is this an answer to our questions? (a) My promise can now be understood, without circularity, as involving my intention to promise. "For it is clear that what you do is not a move in a game unless the game is being played and you are one of the players . . . That involves . . . appropriate *expectations* and *calculations*" in connection with your proceedings (1981a, III: 17). That is, in order to have the requisite intention *of promising to* ϕ, I need not administer an *account* of promising to myself; rather what I *do and think* has to be in line with a certain *practice*. (b) To see how mere words can create real restrictions, we need no more than look at the impossibility, which issues from one's bumping to ϕ, of avoiding the danger of unwelcome consequences unless one ϕ's.

(4) This restriction, however, is not yet (a) moral requirement, or (b) a necessity to respect it. (Hence that paper's full title, "On Promising," is followed by "and Its Justice, and Whether It Need Be Respected *in Foro Interno*.") Is there a (prima-facie) need to keep promises and respect rights – some of them of course created by promises – which goes beyond the necessity internal to the linguistic practice? (a) This practice provides us, *inter alia*, with a way of "getting one another to do things without the application of physical force," and this "is a necessity for human life" (1981a, III: 18) in Aristotle's sense of "that without which good cannot be or come to be" (p. 15). Hence obligation: "a restriction" on "one's possibility of acting well" (p. 15). (b) The practical necessity arising from the common good is not *eo ipso* one from the point of view of my own good. Rather, Anscombe holds, "if someone does genuinely *take* a proof that without doing X he cannot act well as a proof that he must do X, then this shows . . . that he has a purpose that can be served only by acting well, as such" (p. 19). Note that neither (a) nor (b) type necessities are created by convention, though the second is, in a sense, brought about by my intention – an overall purpose in life – unless "man has a last end which governs all" (1995: 34).

(5) Anscombe's chief contribution to political theory and the philosophy of law, "On the Source of the Authority of the State" (1981a, III: 130–55), is partly based on her account of rights. Political government, to be distinguished both "from authority in voluntary co-operative enterprises" and from "control of a place by a gang of bandits," must be characterized "by its authority in the command of violence" (p. 132). Since authority is a right to give orders and make decisions, we may hope to explain it by describing a language-game in which "It is N's right to ϕ" gets its meaning, originally as a prelude only to "So he/she can ϕ," "So you can't ϕ," etc., from a practice of preventing anyone but N from ϕ-ing, of reproaching them for interference with N's ϕ-ing, and so on (cf. (3) above). Given this explanation, N's right to ϕ may yet be merely customary ("conventional"), and perhaps even an injustice. A way of proving that it is not (if it is not), is to show that ϕ-ing is needed for the performance of a *task* which it is *practically necessary* that N perform. For here we have a non-conventional "N must," and it entails the conventional "N can" which ascribes a right to N. An existing government G might then be shown to have political authority – a right to enforce obedience by the threat and use of violence – by showing that (a) such enforcement is needed

for government, (b) (in view of how men tend to treat each other) government is a task necessary for human good, and (c) it is G that customary right or some practical necessity require to govern.

Challenges to contemporary moral philosophy

"In general, my interest in moral philosophy has been more in particular moral questions than in what is now called 'meta-ethics' " (1981a, III: viii). Some of these questions relate to topics in social ethics: parental authority (pp. 43–8, 135); state, law, and punishment (pp. 51–60; 123–55); war (pp. 51–81). Others concern contraception (pp. 82–96), murder (pp. 51–61), and topics in medical ethics. Some of Anscombe's themes are "topic-neutral." They include the ones discussed here on pp. 320–4, the problem of "Authority in Morals" (1981a, III: 43–50), absolute prohibitions, and the anti-consequentialist principle that you are not responsible for foreseeable but unintended consequences of your actions *in the way* you are for chosen means and ends (see pp. 54–5, 58–60, and 78–9).

More widely noticed than her treatment of particular moral questions has been Anscombe's 1958 article "Modern Moral Philosophy," a spirited defense of three theses: (1) We should stop doing moral philosophy "until we have an adequate philosophy of psychology" (1981a, III: 26). For an adequate account of acting well must be based on a philosophical understanding of human nature and such concepts as action, pleasure, need and want, intention, motive, and virtue. (2) "The moral sense of 'ought' " is an illusion due to reminiscences of a "law conception" of ethics which has long since been given up (pp. 26, 29–33). If you do not believe in God as a law-giver (as Stoics, Jews, and Christians do), what remains of an *obligation* to act well is the *word*, spoken with a *special emphasis and feeling* (plus vain attempts to ground the "moral law" in individual autonomy or social contract). Hence " 'morally wrong' *both* goes beyond the mere factual description 'unjust' *and* seems to have no discernible content except a certain compelling force, which I should call purely psychological" (p. 41). Without the assumption of divine legislation, there is indeed no "ought" from "is" that is not of the kind discussed in the previous section: necessity by convention and by the practical requirements of common or individual human good, or "flourishing" (pp. 38–42). (3) The differences between the well-known English moral philosophers since Sidgwick are "of little importance," compared with their common "consequentialism," that is, their rejection of the Hebrew–Christian conviction that "certain things [are] forbidden whatever *consequences* threaten, such as: choosing to kill the innocent for any purpose, however good; vicarious punishment; treachery," etc. (pp. 26, 34–6).

Anscombe's challenges to moral philosophy have been taken up over the last decades by some who conceive of ethics on broadly Aristotelian lines. In other quarters, however, her substantial and critical contributions to this as to other areas of philosophy have not yet received the attention they deserve.

Notes

1 Cf. the more elaborate argument in "Soft Determinism" (1981a, II: 163–72). Cf. also "Chisholm on Action," *Grazer Philosophische Studien* 7/8 (1979), 205–13; and "The

Causation of Action," in *Knowledge and the Mind*, ed. C. Ginet and S. Shoemaker (New York and Oxford: Oxford University Press, 1983), pp. 174–90. Here Anscombe examines the relations between a physiological investigation into the causes of human actions, an account of agency in terms of intentions, and historical explanations. She argues that the second and third are in some sense *supervenient* only, if determinism is true.

2 Because of the common structure of the two contexts, Anscombe here keeps to this spelling, rather than "intensional," reminding us of the etymological background: "intendere arcum in" = "to shoot at." In the philosophy of logic, the topic of intensionality is typically treated in discussions of identity, or modal or belief contexts, within which an intensional object in Anscombe's sense *corresponds* to the sense of, e.g., a name.

Bibliography

Works by Anscombe

1959: *An Introduction to Wittgenstein's* Tractatus, London: Hutchinson. (Reprinted 1971.)

1961 (with Geach, P. T.): *Three Philosophers*, Oxford: Blackwell Publishers.

1963: *Intention*, Oxford: Blackwell Publishers. (First published 1957.)

1979: "Prolegomenon to a Pursuit of the Definition of Murder: The Illegal and the Unlawful," *Dialectics and Humanism* 6, pp. 73–7.

1981a: *Collected Philosophical Papers*, vol. I, *From Parmenides to Wittgenstein*; vol. II, *Metaphysics and the Philosophy of Mind*; vol. III, *Ethics, Religion and Politics*, Minneapolis: University of Minnesota Press.

1981b: "A Theory of Language," in *Perspectives on the Philosophy of Wittgenstein*, ed. I. Block, Oxford: Blackwell Publishers.

1982a: "Medalist's Address: Action, Intention and 'Double Effect'," *Proceedings of the Catholic Philosophical Association* 56, pp. 12–25.

1982b: "On Private Ostensive Definition," in *Language and Ontology: Proceedings of the 6th International Wittgenstein Symposium*, ed. W. Leinfellner, Vienna: Hölder.

1983: "Sins of Omission: The Non-treatment of Controls in Clinical Trials," *Proceedings of the Aristotelian Society*, suppl. vol. 57, pp. 223–7.

1987: *Absicht* (German edn. of *Intention*), ed. J. M. Connolly and T. Keutner, Freiburg and Munich: Alber. (Contains a list of works on Anscombe.)

1989: "Von Wright on Practical Inference," in *The Philosophy of Georg Henrik von Wright*, ed. P. A. Schilpp, La Salle, IL: Open Court.

1995: "Practical Inference," in *Virtues and Reasons: Philippa Foot and Moral Theory*, ed. R. Hursthouse, G. Lawrence, and W. Quinn, Oxford: Clarendon Press.

Works by other authors

Diamond, C. and Teichman, J. (eds.) (1979) *Intention and Intentionality. Essays in Honour of G. E. M. Anscombe*, Brighton: Harvester. (Contains a "Bibliographic note" on Anscombe and a list of works on her.)

Gormally, L. (ed.) (1994) *Moral Truth and Moral Tradition: Essays in Honour of Peter Geach and Elizabeth Anscombe*, Blackrock: Four Courts. (Contains a list of works on Anscombe.)

Thompson, M. (1998) "Anscombe, G. E. M.," in *Routledge Encyclopedia of Philosophy*, vol. I, ed. E. Craig, London: Routledge, pp. 280–3.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

25

R. M. Hare (1919–)

WALTER SINNOTT-ARMSTRONG

Richard Mervyn Hare has written on a wide variety of topics, from Plato to the philosophy of language, religion, and education, as well as on applied ethics, but he is best known for his general moral theory. Hare's views on ethics have developed since his groundbreaking book, *The Language of Morals* (1952), but the main thrust of his position has remained fairly constant.

Definition of moral judgments

Hare defines the class of moral judgments to include any judgment that is prescriptive, universalizable, and overriding (1981: 53–7). Prescriptivity distinguishes moral judgments from judgments of natural science and history. Universalizability separates moral judgments from particular commands, such as by army sergeants, as well as from legal judgments (1963: 36). Overridingness divides moral judgments from aesthetic value judgments (1963: 139).

Hare's definition of moral judgments is formal in that it does not require any particular content. It is possible, on Hare's definition, to make a moral judgment that one should never step on cracks in the sidewalk, even if such steps harm nobody, break no promise or law, and so on. Some critics find this implication unpalatable, but Hare responds that his definition still captures one possible and useful specification of the term "moral." Many people seek to formulate and justify a system of judgments that are moral in Hare's sense.

Prescriptivism

Hare's moral theory, then, starts with prescriptivism, which is the view that moral and other value judgments are typically prescriptive. To call a judgment prescriptive is to say that it is used to prescribe action, that is, to perform some speech act in a large group that includes commanding, advising, encouraging, discouraging, and so on, with respect to some particular action or kind of action. The paradigm form of prescription is the imperative, so prescriptivism claims that a value judgment such as "Hondas are better than Toyotas" are used to perform a speech act similar to "Choose a Honda over a Toyota." Despite the analogies between value judgments and imperatives, Hare has

insisted from the start that "it is no part of [his] purpose to 'reduce' moral language to imperatives" (1952: 2).

Prescriptivism is best understood in contrast with its predecessors in metaethics. G. E. Moore criticized naturalism, which is the view that moral and other value judgments ascribe or deny evaluative properties, such as goodness or rightness, that are supposed to be reducible in some way to properties that are natural apparently in the sense that they can be studied by the methods of natural science. In place of naturalism, Moore proposed non-naturalism, which is the view that moral and other value judgments are about a wholly different kind of property that can be known only by intuition.

Hare accepts Moore's arguments against naturalism and adds new arguments of his own (1952: 79–93). However, Hare also rejects Moore's non-naturalism as unnecessarily mysterious. Moore's mistake, according to Hare, was to retain a basic assumption of naturalism, namely, descriptivism, which is the view that value judgments are used to describe values or evaluative properties. Hare argues that both naturalism and non-naturalism should be rejected, because value judgments are not used to describe in either way.

Descriptivism was rejected before Hare by emotivists, including A. J. Ayer and Charles Stevenson (see AYER and STEVENSON). Emotivists claimed that value judgments are used to express and arouse emotions. Against this view, Hare points out that one can make value judgments dispassionately, without having any feeling or emotion to express. Moreover, just as advice can be given but not followed, so a value judgment can succeed in its speech act even if it does not arouse any emotion or change any behavior in the audience. These points enable Hare to distinguish value judgments on his view from propaganda and to subject them to rational scrutiny of a kind that is often assumed to be inappropriate for emotions (1952: 9–16).

Hare's view, then, is that value judgments are typically used for a different speech act, namely, prescribing. He does admit that some value judgments are not used to prescribe. When someone says a building is good Gothic revival just to indicate that it is the kind of building that would be judged good by people who like Gothic revival buildings, Hare calls this an "inverted commas use" (1952: 124). Value judgments can also be used ironically and merely to pay lip service to conventions. Past-tense value judgments are explained in similar ways. When I say that it was morally wrong for Jefferson to hold slaves, I do not prescribe to Jefferson that he do anything, since it is too late for that; but, if I had been able to do so, then I would have prescribed that he not hold slaves. In such cases, value judgments are not used to prescribe, but such uses are still parasitic on, because explained in terms of, prescriptive uses by other people or in other circumstances. Hare's claim that value judgments are "typically" prescriptive seems to mean that all value judgments either are used to prescribe or can be explained in terms of other uses that are prescriptive (1963: 22n).

Such prescriptivism provides a natural explanation of many features of evaluative language. One common dictum claims that "ought" implies "can." If a judgment that an agent ought to do an act is used to prescribe that the agent do the act, and there is something wrong with prescribing that an agent do an act when the agent cannot do it, then this explains why there is something wrong with saying that an agent ought to do an act that she cannot do (1963: 51–66).

A more often questioned implication of Hare's prescriptivism is that one cannot fully think that one ought to do an act and yet not do it, if one can do it and now is the time to do it. This seems to rule out weakness of will, but Hare explains apparent cases of weakness of will as cases where agents do not really think at the time that they ought to do relevant particular acts, use parasitic senses of "ought," psychologically cannot bring themselves to do what they think they ought to do, and so on (1963: 67–85). If such reinterpretations can adequately explain away all apparent cases of weakness of will, then Hare's prescriptivism is harder to refute than his critics assume.

Universalizability

Despite their prescriptivity, value judgments differ from singular imperatives in a crucial respect, according to Hare. When a drill sergeant commands, "To the left, march!," he needs no reason for choosing left over right, and nothing goes wrong if, when the same circumstances arise again later, he then commands, "To the right, march!" In contrast, Hare argues that there is a logical inconsistency between saying that one agent ought to do an act and denying that another agent ought to do an act with relevantly similar properties in relevantly similar circumstances (1952: 81). Hare's explanation of this linguistic rule is that "all value judgments are covertly universal in character, which is the same as to say that they refer to, and express acceptance of, a standard which has application to other similar instances" (1952: 129), that is, they presuppose a general principle. This implicit universality makes it legitimate to ask for a reason why the agent ought to do the act, or a property that makes the thing good. It also enables value judgments to be useful for public teaching of standards (1952: 134). In this respect, value judgments are like descriptive judgments, according to Hare (1963: 10–14).

This doctrine of universalizability might seem empty without limits on which properties are relevant. Hare does claim that references to individuals, as in proper names or indexicals, must be irrelevant in order for the presupposed standard or principle to be universal. However, he insists that no additional limits are part of the shared meaning of evaluative terms. In his view, we cannot determine which properties are morally relevant in advance of determining which moral principles are defensible (1981: 62–4). This makes Hare's doctrine of universalizability much weaker than that of his predecessor Kant on most interpretations.

Nonetheless, Hare's thesis of universalizability is criticized by particularists. Some argue that even an agent's individual identity or particular spatiotemporal location might be morally relevant; or at least that this is not excluded by language alone, so to assume otherwise is to adopt a substantive position. Other particularists argue that the reasons why one act is morally wrong might not make another act morally wrong, because the force of each morally relevant feature varies with the circumstances in ways that cannot be specified in any general principles. Hare would respond by arguing that some feature of the person or time or circumstances must be available to explain why one act is wrong when another is not. We might and need not be able in practice to specify fully the relevant properties or underlying principles, but some specification must be possible in theory for our value judgments to be logically consistent, in Hare's view (1963: 18–20).

Rationality

The next prong of Hare's theory is a particular view of rationality. The linguistic theses of prescriptivism and universalizability lay down some limits on consistency and, hence, rationality. Hare also assumes that "any rational thinking about [moral questions] has to be done in the light of facts" (1981: 87). At times, Hare also seems to assume that logical consistency and knowledge of facts is all there is to rationality.

This account of rationality might not seem very controversial, but Hare adds that one fact that needs to be considered for moral judgments to be rational is "what it is like" for people who are affected by our actions (1981: 92). In particular, if they suffer, it is not enough for me to know that they suffer. I need to know what their suffering is like to them.

This kind of knowledge, Hare argues, requires me to have a certain motivation or preference. The following two claims are distinct for any situation:

1 I now prefer with strength *S* that if I were in that situation *x* should happen rather than not.
2 If I were in that situation, I would prefer with strength *S* that *x* should happen rather than not. (1981: 95)

(1) is about my current preferences regarding a counterfactual situation, whereas (2) is about a counterfactual situation in which I would have preferences that I do not currently have. Although Hare admits that these claims are distinct, he argues that "I cannot know that (2), and what that would be like, without (1) being true, and . . . this is a conceptual truth, in the sense of 'know' that moral thinking demands" (1981: 96). This transfer principle, as I will call it, implies that one must have certain preferences in order to be rational.

The master argument

By combining prescriptivism, universalizability, and rationality, Hare claims to be able to derive a kind of utilitarianism (1981: 109–11). The argument starts with an example, whose facts I have modified somewhat to clarify the logic of the argument:

1 B cannot park B's car in this parking place unless I move my bicycle.
2 B has a strong preference to park B's car in this parking place.
3 I have only a weaker preference not to move my bicycle (or to leave it there).

Utilitarianism suggests that I morally ought to move my bike, so Hare's goal is to derive an absurdity from the supposition that an opposite belief is rational:

4 I rationally believe that I morally ought to leave my bicycle there.

This supposition on Hare's view of rationality implies

5 I know all of the relevant facts about leaving my bicycle there, including what it would be like for B if I left my bicycle there.

Assuming that roles include preferences, so that, when we switch roles, we also switch preferences, (2) and (5) imply

6 I know that, if B and I switched roles, I would strongly prefer that B not leave the bicycle there.

Hare's transfer principle plus (6) yields

7 I now strongly prefer that, if B and I switched roles, B would not leave the bicycle there.

Assuming that this is the strongest preference between these alternatives in this situation, then, since preferences are prescriptive,

8 I accept the prescription: "if B and I switched roles, let B not leave the bicycle there."

But the universalizability of moral judgments plus (4) imply

9 I believe that, if B and I switched roles, B morally ought to leave the bicycle there.

This plus the prescriptivity of moral judgments implies

10 I accept the prescription: "if B and I switched roles, let B leave the bicycle there."

The prescriptions in (8) and (10) are inconsistent, and inconsistency excludes rationality, so (4) is refuted. This means that

11 If I am rational, I cannot believe that I morally ought to leave the bicycle there.

Once this belief is ruled out as irrational, the options are limited:

12 If I believe that I morally either ought or ought not to leave the bicycle there, and if I am rational, then I must believe that I morally ought not to leave the bicycle there.

Assuming that this example is not different in any relevant respect from any other conflict involving only two people, we can generalize:

13 In every two-person conflict, if I believe that I morally either ought or ought not to an act, and if I am rational, then I must believe that I morally ought not to do whatever frustrates stronger preferences than it satisfies.

What is believed in the consequent of (13) is just what preference utilitarianism says about such two-person conflicts.

Hare next extends his argument to choices that affect any number of people. Then, to make a rational choice, an agent needs to know what it is like for each person who is affected. That knowledge, according to Hare's transfer principle, requires the agent to have current preferences for or against the act being done if he were to switch roles with each of them and take on their positive or negative preferences. Indeed, the agent must have preferences with the same strengths as those of the people affected. Thus, an agent morally ought to do what he would prefer to be done if he had to occupy successively each of the roles of each person affected. This conclusion is a general version of preference utilitarianism.

Hare's argument could be criticized from many angles. First, amoralists might refuse to make any positive moral judgments at all. Then they do not deny (12) or (13), but

they do block any derivation of their consequents. Hare admits that there is nothing "logically inconsistent in this position" (1981: 186), but he gives "reasons of a non-moral sort why it should not be chosen" (p. 190).

A different problem is raised by fanatics, who cling to ideals even when they know that their moral judgments run contrary to utilitarianism. Hare responds that the fanatic's moral judgments are in effect a new moralistic kind of preference, so fanatics face a dilemma (1981: 181). On one horn of this dilemma, the fanatic claims that his moralistic preferences are strong enough to outweigh all of the preferences of everyone else affected by the action. This claim would be implausible, and, even if it were true, a preference utilitarian could just grant that this moralistic preference should be fulfilled, since it is stronger than the rest put together. Thus, if the fanatic's position is to remain contrary to utilitarianism, the fanatic must grab another horn of Hare's dilemma. The only alternative for the fanatic is to admit that his moralistic preference is not strong enough to outweigh all of the preferences of everyone else affected by the action. But then, if the fanatic is rational, he must know what it is like for those people's preferences to be frustrated. By the transfer principle, he must then come to have corresponding preferences with the same strengths. These preferences will then make him abandon his fanaticism. What he cannot do is remain a fanatic and also remain rational, according to Hare.

The transfer principle itself faces problems, however. Consider someone who is trying to kill me so that they can take my place on the basketball team. Should I stop them? For my answer to be rational, I need to know that, if we switched places (so I wanted to kill them in order to get on the team), then I would strongly prefer that they not stop me from killing them. According to the transfer principle, I cannot know what that would be like unless I *now* strongly prefer that if I tried to kill them to get on the team, then they would not stop me. However, this current preference does not seem necessary for me to know all that is going on. I can know why they want to kill me, and what it is like for them, even if my current moral beliefs keep me from having any current preference that they not stop me if I try to kill them. More generally, preferences that we see as immoral seem to be one kind of preference that rationality does not require us to take over from others in order to know what it is like for them to have that preference frustrated.

The same problem arises when the transfer principle is used against non-utilitarian moral theories. Suppose a retributivist believes that the government morally ought to execute a convicted murderer, even if this execution would not maximize preference satisfaction. This position is contrary to utilitarianism, so Hare would label this retributivist a fanatic. Whatever he is called, such a retributivist can know that, if he were in the convicted murderer's situation, then he would prefer strongly that the execution not happen. He can even know what it would be like to be executed, as much as anybody else can. But the retributivist still need not *currently* have any preference that, if he were in the same situation as the convicted murderer, he should not be executed. In that counterfactual situation, he himself would deserve to be executed, according to his current retributivist views; so, if he sticks with his views, he would prefer that, if he himself had committed such a heinous murder, then he himself should be punished like any other murderer. Such retributivism seems conceptually coherent, consistent with both prescriptivism and universalizability, and also common. Thus, Hare's

transfer principle seems to fail in the very kind of case where he needs it. If so, his argument cannot prove his version of preference utilitarianism.

Utilitarianism

Even if Hare's argument does not prove his version of utilitarianism, that conclusion still might be correct and defensible. Hare is basically an act utilitarian with a preference-based theory of value. Such views have been subjected to a great deal of criticism and even scorn, but Hare's version has distinctive features that make it harder to refute than many others.

The most important innovation is Hare's distinction between two levels of moral thinking. This distinction is motivated by a discussion of apparent moral conflicts. In his main example (1981: 27), a lifelong friend visits unexpectedly from Australia and asks to be shown around Oxford on the very day when Hare promised to take his children for a picnic on the river. It seems initially that Hare morally ought to keep his promise, and he also morally ought to show his friend around Oxford. The problem is that Hare cannot do both, so the fact that he morally ought to keep his promise implies that he morally ought not to show his friend around Oxford. Since moral judgments are prescriptive and overriding, so "ought" implies "can," Hare denies the possibility that he morally both ought and ought not to show his friend around Oxford. So he also denies that such moral conflicts are ever possible in the end.

Nonetheless, Hare wants to explain why moral conflicts seem to be not only possible but common. To explain this appearance, Hare distinguishes two uses of "ought" that occur at two levels of moral thinking. On the intuitive level, Hare's thinking that he morally ought to show his friend around Oxford is "inseparable from" his feeling compunction before and guilt after his failure to show his friend around Oxford (1981: 30–1). Since a good person would also feel compunction before and guilt after failing to keep his promise to his children, Hare admits that conflicting moral beliefs at the intuitive level can both be justified by their practical usefulness. Intuitive moral beliefs facilitate moral teaching and reduce errors in moral judgment when information is scarce or time is short. That explains why we do and should sometimes think at the intuitive level, where moral conflicts seem to occur.

But this intuitive level of moral thinking is not enough by itself. We need to determine which intuitive moral principles should be taught and used in making everyday decisions. We also need some way to determine the morally right course of action when intuitive moral judgments conflict. Both of these problems, according to Hare, can be solved only at a different level of moral thinking, which he calls the critical level. At that critical level, to say that Hare morally ought to show his friend around Oxford is to make a prescriptive and overriding judgment, so such critical moral judgments cannot conflict. Moreover, critical moral thinking must conform to preference utilitarianism, according to Hare's master argument. Since incompatible alternatives cannot each frustrate stronger preferences than it satisfies, Hare's utilitarianism implies that an agent never morally ought to adopt each of two conflicting options. Such critical thinking in conformity with utilitarianism is what ultimately determines which act is morally right, according to Hare. That is why Hare can remain a utilitarian and yet still explain why moral conflicts do and should seem to occur.

This dichotomy between levels of moral thinking also enables Hare to respond to standard counterexamples that are supposed to refute utilitarianism by showing that it conflicts with common moral intuition: "Perhaps the [proverbial] sheriff should hang the innocent man in order to prevent the riot in which there will be many deaths, if he knows that the man's innocence will never be discovered and that the bad indirect effects will not outweigh the good direct effects; but in practice he will never know this" (1981: 164). Hare's point is that, in order to create any trouble for utilitarianism, the details of such examples must be so unrealistic that we have little or no reason to trust our moral intuitions about such circumstances. Our moral intuitions are justified because we will act better in most situations if we inculcate such moral intuitions deeply into our characters, but the same moral intuitions cease to be reliable guides in unrealistic circumstances for which they were never intended. Consequently, utilitarianism cannot be refuted by any appeal to moral intuition in this or any other unrealistic example, according to Hare (1981: 134).

This response will not convince every opponent, and many other objections could be raised to Hare's utilitarianism and to other aspects of his moral theory. Nonetheless, the range of Hare's views, his attention to detail, the clarity of his writing, and his ability to bring abstract moral theory to bear on concrete issues of great importance provide a model for all moral philosophers to follow.

Bibliography

Works by Hare

1952: *The Language of Morals*, Oxford: Clarendon Press.
1963: *Freedom and Reason*, New York: Oxford University Press.
1971a: *Essays on Philosophical Method*, London: Macmillan.
1971b: *Practical Inferences*, London: Macmillan.
1972a: *Applications of Moral Philosophy*, London: Macmillan.
1972b: *Essays on the Moral Concepts*, London: Macmillan.
1981: *Moral Thinking: Its Levels, Method, and Point*, New York and Oxford: Oxford University Press.
1982: *Plato*, New York and Oxford: Oxford University Press.
1989a: *Essays in Ethical Theory*, New York and Oxford: Oxford University Press.
1989b: *Essays on Political Morality*, New York and Oxford: Oxford University Press.
1992: *Essays on Religion and Education*, New York and Oxford: Oxford University Press.
1993: *Essays on Bioethics*, New York and Oxford: Oxford University Press.
1998: *Sorting Out Ethics*, New York and Oxford: Oxford University Press.
1999: *Objective Prescriptions and Other Essays*, New York and Oxford: Oxford University Press.

Work by other authors

Seanor, D. and Fotion, N. (eds.) (1990) *Hare and Critics: Essays on Moral Thinking*, Oxford: Clarendon Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

26

P. F. Strawson (1919–)

P. F. SNOWDON

Life

Peter Frederick Strawson was educated at St. John's College, Oxford, where he read Philosophy, Politics, and Economics, graduating in 1940. He then served for six years in the British Army, becoming a captain. After a short period as a lecturer at the University of North Wales, Bangor, he returned to Oxford, becoming a Fellow at University College in 1948. In 1968 Strawson was appointed Gilbert Ryle's successor as Waynflete Professor of Metaphysical Philosophy. He was made a Fellow of the British Academy in 1960, was knighted in 1977, and retired in 1987, though since then he has continued his philosophical activities. Strawson's major publications include *Individuals* (1959), an exploration in what he called "descriptive metaphysics," *The Bounds of Sense* (1966), a constructive and critical study of Kant, *Scepticism and Naturalism* (1985), a study of both general skepticism and some more specific variants, and *Logico-Linguistic Papers* (1971), a collection of his highly influential papers about language, including "On Referring," the article that first made Strawson famous.

Themes

Strawson established his pre-eminence in postwar Oxford philosophy by the extraordinary range and depth of his work. He has written about the philosophy of language, of logic, metaphysics, epistemology, the history of philosophy, but also about the nature of philosophy itself. And within each of these broad areas he has investigated many topics. Thus, within the philosophy of language he has written about reference, meaning, truth, the subject/predicate distinction, speech acts, the meaning of connectives, and the nature of grammar. One aspect of the depth of Strawson's work has been his attempt to establish explanatory links between the different branches of philosophy. For example, he illuminatingly links the metaphysical distinction between particular and universal to the logical subject/predicate distinction. The special quality of his work resides in his ability to develop original ideas across such a wide range, the care and ingenuity with which he develops these ideas, and a persistent tendency to pursue issues to a deep level. Strawson also writes in a stylish, distinctive, and untechni-

cal manner, conferring on his works an elegance as literature unusual for recent philosophy.

Strawson's views have developed and been modified, and he has neither tended to repeat himself nor engaged overly much with the extensive critical discussions of his work. In consequence, and in contrast with some other leading philosophers, there is no core set of repeatedly defended doctrines which might be called Strawson's philosophy. There are, however, certain abiding and recurring themes in his writing which deserve to be formulated.

Strawson's picture of human thought, which has links, in different ways, to those of Hume and Kant, is that, despite its impressive development over time, with, for example, the emergence of science, and the improvement of its understanding in all domains, there is an abiding, fundamental, unrevisable framework. As he says in the Introduction to *Individuals*,

> there is a massive central core of human thinking which has no history – or none recorded in histories of thought; there are categories and concepts which, in their most fundamental character, change not at all. Obviously these are not the specialities of the most refined thinking. They are the commonplaces of the least refined thinking; and are yet the indispensable core of the conceptual equipment of the most sophisticated human beings. (1959: 10)

Charting these central concepts is one task for philosophy, the task which Strawson calls descriptive metaphysics. This task is theoretical and constructive, and represents a vision of philosophy which contrasts significantly with what was a more negative, critical, and piecemeal approach associated with Austin, an approach which dominated part of the Oxford landscape in the early stage of Strawson's presence there. Strawson's development of his own program significantly changed that landscape.

According to Strawson, one part of this framework is the physical world of perceptible bodies, constituting an abiding framework in space and time, with their manifest, and not so manifest, causal properties; another part is persons, entities with both bodily attributes, such as weight and height, and also psychological attributes, such as consciousness, perception, thought, and action, and, of course, an understanding of the very concepts being described. Persons understand and use language which is shaped to express the basic concepts. Our employment of these categories cannot be eliminated in favor of those of science, since the categories of science itself are accessible only via the basic framework. Further, these categories do not earn their right to employment by being justified in the light of arguments based on experience as traditionally conceived by empiricists, for there is no such neutral experience describable in more basic terms. The rejection of the empiricist conception of experience, advanced in a relatively sophisticated form by A. J. Ayer, represents one theme shared by both Austin and Strawson (see AUSTIN; cf. AYER). These categories do not earn their right to employment, either, by being shown to be reducible to more basic categories, for no such reduction is possible or needed. As an ontologist, therefore, we might call Strawson's attitude "relaxed realism." He endorses the reality of such entities and their properties – hence, the realism – without supposing that this requires some strong unification between the different levels of thing or property – hence the relaxedness.

Further, although the grounds presented for saying this have changed, Strawson regards skeptical criticisms of the framework as essentially based on misunderstandings of one form or another, and Strawson's main (although certainly not sole) epistemological interest has been to display the errors of skepticism. In his early discussion of induction, and also in *Individuals*, he appears to suggest that the skeptical thought that in circumstance C (the best circumstances that can obtain) we do not know that P (or do not reasonably believe that P) can be shown to be inconsistent with the meaning of P. In *The Bounds of Sense* Strawson sympathetically explores a strategy of transcendental arguments, according to which there is an inconsistency in the skeptic's attitude, in that the concepts the skeptic is prepared to apply presuppose the application of the concepts about which he is skeptical. More recently Strawson has attempted to discredit skepticism on the basis of its inability to genuinely persuade anyone. As well as such extreme traditional skepticism, Strawson is also opposed to the more limited skepticism of some current philosophers, such as Quine, who reject parts of our conceptual scheme, in his case those to do with meaning or psychological states. Strawson argues that there are no legitimate metaphysical requirements which such notions fail to satisfy, and they are, moreover, indispensable to our thought and to the very inquiries which are supposed to supersede them. There is, in Strawson's approach, a conceptual conservatism similar to Wittgenstein's, but Strawson reveals no sympathy with Wittgenstein's opposition to theoretical philosophy.

One fundamental task of philosophy is to describe this framework, to display connections between the basic categories, (say between the categories of perception and causation), to disarm philosophical skepticisms or reductions, and to describe, in a realistic way, the language we have for expressing these concepts, without restricting the categories we employ in the description of our language to those favored by formal logicians, whose purposes are rather different.

Definite descriptions and reference

Strawson's first major publication was "On Referring," which, in 1950, established for him more or less immediately, an international reputation, and also initiated a still continuing debate about the nature of reference. In that article, Strawson's principal aim was to criticize and replace Russell's famous theory of definite descriptions. Noun-phrases beginning with the definite article, for example, "The Prime Minister of England," "The man over there," "The cleverest man in the world," are called definite descriptions. Russell's theory proposes that a sentence of the form "The F is G" (call this sentence S), is equivalent to, or is to be analyzed as, "There is one and only one F and it is G" (call this sentence SR). According to this analysis the occurrence of a definite description signals the assertion of an existential claim, "there is an F," and a uniqueness claim, "and only one F." The analysis also implies that the utterance of sentence S is false if there is no such thing as "the F," since S is analyzed as saying that there is an F (see RUSSELL).

Strawson makes three main criticisms. (1) He first argues that the theory is unsupported. According to Strawson, Russell regarded S as equivalent to SR because he thought that sentences of the S form are meaningful even if there is no F, which they could not be if they are of subject-predicate form. Strawson comments that it is neces-

sary in thinking about language to distinguish between roughly a sentence and the use of a sentence. Sentences have meaning, but that does not require that each use of a meaningful sentence expresses a true-or-false assertion. Strawson's distinction is important and has had a prominent place in recent theories of indexicals and demonstratives. However, the claim that Russell's overlooking of the distinction is "the source of Russell's mistake" has not been widely accepted. The reason is that Russell's own arguments and intuitions primarily relate to what Strawson calls "uses." It is the content conveyed by uses to which Russell is attending. (2) Strawson further argued that S would not be regarded as false in cases where there is no F, rather the question of its truth or falsity does not arise. This became known as the "truth-value gap" thesis. (3) Strawson's main argument, however, is that it is obvious that a speaker who uses "the F" is simply not *saying* that there exists such a thing as the F; rather the speaker implies that there is an F by employing "the F," in speaking to the audience, to refer to the object. In this respect he compared the use of definite descriptions to that of demonstratives (such as "that" and "this"). Arguments (2) and (3) were both contested and debated. The existence of truth-value gaps was denied, and it was also denied that we can just tell what we are saying in such cases. Further, evidence was produced that supported the conclusion that in many cases "the" is not a device of reference, for example, as in the sentence "The person that each man most admires is his mother."

In subsequent work Strawson refined and limited his account. In *Introduction to Logical Theory* he introduced the term "presupposition" for the relation that he thought existed between saying "The F is G" and the claim "There is an F." Roughly, presupposition holds between P and Q if the truth of P requires the truth of Q, but the falsity of Q does not require the falsity of P. If Q is false the question of P's truth does not arise. This terminology and the investigation of such a relation (or related relations) has been more prominent in linguistics than philosophy.

Strawson added two other important ideas. First he investigated the nature of reference (both in *Individuals* and "Identifying Reference and Truth Values" in Strawson 1971) and provided an account or model of what he calls "identifying reference." The very commonsensical idea is that both speaker and audience have their respective and different knowledge of objects in the world, and in the central case the speaker chooses a referring expression that he judges appropriate to enable the audience to identify amongst the objects they know, the one being spoken of. The speaker invokes or rather relies on the audiences's knowledge of the object but does not need to inform or tell the audience of the object's existence. This is an amplification of Strawson's central intuition about the use of the definite article. Second, Strawson persuasively separated the claim that at least one role of definite descriptions is to make identifying reference from the claim that sentences containing empty definite descriptions are neither true nor false (the truth-value gap thesis). He claims that identifying reference can be characterized without implying that such gaps exist. He proposed instead that the consequence for truth-value of reference-failure is partly determined by the relation between the definite description and the different topics of the discourse. Where the description aims at fixing the topic the result of reference failure is a truth-value gap; where the description figures in a supplementing claim about another topic, the result is falsehood. For example, if, in a talk on the constitution of England, I start by telling you about the president of England, you would dismiss my remarks as confused, but if, in

describing the visitors to the Tower of London I listed the president of England you would dismiss that as false.

This proposal has not persuaded everyone and the debate continues, fueled by later important contributions by Keith Donnellan and Kripke. One fundamental question about Strawson's more recent account is whether it accords an over-central role to the notion of identification in understanding reference. It is a mark, though, of the importance of Strawson's contribution that his original article is still influential.

Truth

In an early paper, Strawson endorsed a Ramsey style redundancy theory of truth, according to which the fundamental characterization of truth is that to say it is true that *p* is simply to say that *p* (see RAMSEY). To this he added some observations about the speech acts standardly performed by use of "true," stressing these to such an extent that he was interpreted as endorsing an analysis of "true" solely in terms of the speech acts it is used to perform, a so-called "performative" theory of truth. In the 1950s and 1960s the idea of a performative analysis, by then renounced by Strawson, was ignored and replaced by a debate, involving Strawson, Austin, G. J. Warnock, and others, into the respective merits of the redundancy theory compared to a version of the correspondence theory suggested by Austin, and refined by Warnock. Strawson's approach was to propound and defend against criticism the redundancy theory, and to criticize the correspondence theory. This was the main debate about truth until Dummett, with his anti-realist approach, and Davidson, moved it in new directions (see DUMMETT).

Austin offered the following analysis of truth: a statement is said to be true when the historic state of affairs to which it is correlated by the demonstrative conventions . . . is of a type with which the sentence used in making it is correlated by the descriptive conventions. His aim was to analyze truth as a correspondence relation between statement and world without explaining the correspondence relation, as correspondence theorists have often done, in terms of a structural isomorphism between world and representation. Austin's reference to conventions is meant to avoid such a notion.

Strawson's critical response to Austin is very rich, but we can distinguish three main lines of criticisms. (1) Many attributions of truth cannot be regarded as saying anything about actual statements, as Austin's account seems to imply they have to be. Someone might, for example, begin a talk by saying, "Although it is true that *p*, *q*," without it being necessary that there is a statement that *p* by someone else to be talked about. (2) Strawson treats talk of states of affairs as equivalent to talk of facts, for which he proposes, in effect, its own redundancy theory. Talk of facts cannot figure in a serious analysis of truth, since to say that it is a fact that *p* is equivalent to saying that it is true that *p*, both simply saying *p*. As Strawson puts it; "There is no nuance, except of style, between 'That's true' and 'That's a fact' " (1971: 196). (3) Strawson's major criticism is that "although we use the word 'true' when the semantic conditions described by Austin are fulfilled," the word "true" patently does not state that those conditions are fulfilled. In using the word "true," it is, according to Strawson, obvious that nothing is being *said* about the conventions of language. "It is true that *p*" is no more about language than is "*p*." This argument, the initial concession in which seems not to be entirely consistent with objection (2), resembles Strawson's main argument against

Russell's theory of definite descriptions. In both cases, Strawson is relying on his sense of what is being said or spoken about in particular parts of natural language.

Strawson's criticisms were generally taken as persuasive, but the debate continued in at least two very interesting directions. One arose out of an observation by Warnock that, even if Strawson's criticism (1) had revealed problems for Austin's account, it remains plausible to claim that in ascribing truth very often something is said about a statement. Indeed, Strawson himself had endorsed the common claim that statements rather than sentences are, as it is said, "the bearers of truth." What is it, then, that they bear? Agreeing to this is, though, difficult for the redundancy theory since it does not treat ". . . is true" as expressing a property of anything. Strawson clarified this issue and ingeniously showed that even in a redundancy theory analysis it is possible to include reference to statements. Thus, "S's claim that *p* is true" can be treated as "As S claimed, *p*." It is difficult not to feel that in this debate the real problem which Warnock was gesturing at was lost; the intuition (whether right or wrong) is not simply that a statement is referred to, but that something is ascribed to it, that it is the bearer of something.

Second, Strawson revised his earlier view that Austin's account of the two types of conventions is at least an accurate specification of *when* we use "true." Strawson, in effect, argues that once the referential conventions attaching to certain words in a sentence and the descriptive conventions attaching to others are worked through to determine what is said, there is no discerning separable demonstrative convention attaching to the sentence as a whole to contrast with the descriptive conventions also governing the sentence as a whole. It is a measure of Strawson's success as a critic that Austin's version of the correspondence theory lacks current supporters. Moreover, in the course of his articles Strawson contributed much to the amplification of a redundancy view. (Searle's contribution to Hahn 1998, plus Strawson's reply, illuminate the debate.)

Logical theory

Some account must be given of Strawson's first book, *Introduction to Logical Theory* (1952), but of all of his books it is the one that has dated most and so I shall be brief. The book has three main aims: first, to be an introductory description of formal logic; second to provide a philosophically adequate analysis of the concepts central to thinking about logic, in particular the concept of entailment; and, third, to determine how far the devices of artificial formal logic provide an accurate account of the significance of the expressions in natural language. This last task is simply a generalized form of what is at stake in Strawson's response to Russell.

The first task is elegantly done in many respects, for example in his discussion, in chapter 2, of logical form, but there is no serious attention to the role in formal logic of proof systems, nor is a rigorous semantics developed; and because of this the fundamental contrast between syntactic and semantic notions is not explained, nor are the notions of consistency and completeness. Strawson (in chapter 1, part 3) analyzes the proposition that A entails B as saying that the proposition (A and not B) is self-contradictory, and adds that the defect of self-contradiction is that one does not say anything by uttering a contradiction. However, no clear account is given of self-contradiction, nor is any proper defence given of the claim that self-contradictory sentences

say nothing (rather than that they say something contradictory). It is in relation to the third task that the book is still relevant. In chapter 3 Strawson undertakes a careful comparison between the significance of the formal logical constants and their natural language analogues. He argues that there are significant differences in each case. For example, he claims that "&" is purely conjunctive, whereas "and" can sometimes convey information about temporal order. He also argues that "P → Q" is true if P is false, but that "If P then Q" is not automatically true in those circumstances. Strawson's arguments are ingenious but they stimulated Grice to devise his own theory about how to distinguish what is literally meant and what is otherwise conveyed or implied. In the light of Grice's theory some of Strawson's points look disputable. However, Strawson and others have themselves disputed elements in Grice's theory, and the debate, in particular about conditionals, remains open.

The general slogan that Strawson endorses is that "ordinary expressions have no exact and systematic logic" (1952: 57). Strawson means by this that it is not possible to give to natural language expressions an abstract meaning assignment which exhausts what they count as conveying across all contexts. Strawson's slogan anticipates recent and very fruitful ideas about language. However, Strawson himself did not embed this intuition in a full theory. The final chapter of *Introduction to Logical Theory* is a famous discussion about induction, which I shall consider when discussing Strawson's epistemology.

Meaning and related notions

Only a brief account of Strawson's discussion of meaning and related notions is possible. He has been critical of at least three approaches to meaning in recent philosophy. Of the approach associated with Quine, which is broadly skeptical about a range of intuitive notions of meaning, Strawson has written in numerous places. Against it he makes the following points. First, the skepticism is grounded on arguments which claim that the meaning-notion cannot be adequately explained in certain preferred terms, say behavioral ones, but there is no reason to ground the semantic notions that way. Second, the notions can be validated by the plain agreement between people in the judgments they make. Third, the notions are indispensable to us as language users; thus, we simply cannot speak and think in the way we do without talking of meaning and sameness of meaning. And fourth, the notions are indispensable to the theoretical study of language and logic as well; thus logic cannot do without propositions.

In "Meaning and Truth" (in Strawson 1971) Strawson criticized another approach, that of Davidson, according to which a theory of meaning should be a truth definition (see DAVIDSON). Strawson's argument, which is rather complex, is, in effect, that the notion of truth is secondary to the notion of saying and thinking, and that therefore, meaning, together with truth, has to be grounded in a relation between sentences and the cognitive roles and communicative purposes of speakers and hearers. This idea led him to endorse a modified version of Grice's approach to meaning. In other places he has criticized specific proposals by Davidson about the analysis of language. Finally, Strawson has rejected the anti-realist ideas of Dummett and others, with their approaches to meaning (see DUMMETT). Strawson sees anti-realism as a version of revisionary metaphysics, which is unbelievable and unsupported, and it cannot form the

core of a satisfactory account of the meaning of ordinary judgments. Whether finally successful or not, Strawson's clear and elegant discussions of these views have been influential.

Individuals

Individuals was published in 1959. It proved immediately, and has remained, both controversial and extremely influential. The book is divided into two parts, and it is the first part, called "Particulars," which includes the three very famous chapters, "Bodies," "Sounds," and "Persons," that has attracted special attention, and that I shall describe.

In the Introduction, Strawson sets out the important distinction between descriptive and revisionary metaphysics. "Descriptive metaphysics is content to describe the actual structure of our thought about the world, revisionary metaphysics is concerned to produce a better structure" (p. 9). Strawson makes a number of claims employing this distinction; (1) *Individuals* is an example of descriptive metaphysics; (2) revisionary metaphysics is "at the service of" descriptive metaphysics; (3) Aristotle and Kant are descriptive metaphysicians, Descartes, Leibniz, and Berkeley are revisionary; and (4) descriptive metaphysics being general cannot avail itself solely of the resources generated by ordinary conceptual analysis. This vividly expressed distinction, which is certainly valuable for philosophical taxonomy, and these claims, have often been accepted, but queries can be raised, of which I shall mention two. First, Strawson's own practice is not purely descriptive. He offers explanations, propounds necessities, and rejects criticisms, as well as simply describing. Second, and relatedly, the division does not exhaust the types of metaphysics. There is also what we might call "anti-revisionary metaphysics" which amounts to a defence of the extant conceptual scheme, or a criticism of a suggested revision. As well as being descriptive, Strawson's practice is also anti-revisionary.

In the first chapter, "Bodies," Strawson introduces some important concepts and argues that material bodies are the "basic particulars from the point of view of identification" (p. 5). The first concept is that of a speaker identifying a particular object for an audience. This occurs when a speaker refers to an object and the audience is able to identify the object being referred to. The second concept is that of identification dependence. Thus, it might be that our ability to identify, in the first sense, one sort of particular depends on our ability to identify another sort of particular, but not vice versa. If so, there is identification dependence of the former sort on the latter sort. The third concept is that of reidentification; this is Strawson's term for an identity judgment in which an item encountered on one occasion is identified with an item encountered on another.

Armed with these concepts, Strawson advances three main claims. The first concerns the way we are able, in the course of understanding reference, to identify what items are referred to. Strawson's picture is that we can do so either by locating them amongst those sensibly present to us, roughly, perceivable by us at that time, or by possessing identifying descriptions which items satisfy. Seeing such a two-fold structure to reference is well known. Russell spoke likewise of knowledge by acquaintance and knowledge by description. Strawson's view of perception is, though, quite different from Russell's, and so the items discernible through experience can really be in space. There

is, though, a worry about descriptions: how do we know that only one item falls under them? Strawson's answer is that the particulars we identify are locatable uniquely in the spatiotemporal framework, say at the unique intersection of various spatial coordinates, and we can therefore guarantee uniqueness within that framework. We can relate the general framework to the segment of the world we currently perceive.

Second, Strawson suggests that the intelligibility of locating items in the spatiotemporal framework (e.g. as the fountain in Trafalgar Square) requires relatively abiding structures of reidentifiable items (e.g. the National Gallery). Hence, to talk of objects within this framework is inconsistent with skepticism about reidentification.

Third, Strawson argues that our ability to identify bodies does not depend on an ability to identify particulars of any other kind, but all other kind of particulars, for example, private particulars, such as the pain in my left foot, unobservable particulars, and particular events, do depend on the identification of bodies. "Material bodies, therefore, are basic to particular identification" (1959: 55).

Many features of Strawson's argument have been challenged, but I shall note only two things. First, Strawson argues in the case of some other candidates, notably events, that they are not basic because they do not present a regular enough framework to be the basis for defining a coordinate system. Clearly, this failure (if it is a failure) is contingent. So, the conclusion about the unique status of bodies is itself contingent. Second, Strawson's notion of identification is an interpersonal one; it concerns what a hearer is able to identify as being referred to by another. It must, therefore, be recognized that the dependency thesis need not be true of the individual thought capacities of a single person.

Strawson continues his investigation of the role of space in our conceptual scheme, argued to be fundamental in chapter 1, by seeing to what extent we can imagine a subject of non-spatial experience who is capable of applying concepts of objective and reidentifiable particulars. The contrast that must be present in this conceptual scheme is that between type identity – being the same sort again – and numerical identity – being the same individual again. Strawson chooses a creature with pure sound experience, which he claims would be non-spatial. He asks whether (1) a subject of such experiences could make sense of numerical identity and (2) whether it could make sense of the self/non-self distinction, which Strawson takes to be central to thinking of objects. In considering (1) Strawson's idea is that an analogue of space is necessary. He imaginatively proposes to generate that by putting into the experience a master sound of constant timbre, but varying pitch and loudness, changes in which are meant to represent movement, along with a relatively constant correlation between points on the master sound and collections of other sound to generate the idea of reidentifiable particular objective sounds. Strawson does not claim that this is sufficient, but only that it is not obviously insufficient. In relation to (2) Strawson suggests that there are no very hopeful grounds for introducing the distinction in such an impoverished experience world.

Strawson's imaginative exercise is brilliantly discussed by G. Evans (in Van Straaten 1980), who proposes that Strawson's employment of the master sound underestimates the significance of space and space-occupation in our thought about the world. It can be said, though, that Evans's further insights rest on Strawson's pioneering explorations.

Persons and states of mind

Chapter 3 of *Individuals* is, perhaps, the most discussed chapter of the book, and it deserves a separate section, bringing in, as well, Strawson's later consideration of related issues in *Scepticism and Naturalism*, chapter 3. Strawson's argument begins by picking up on a theme that had emerged in relation to the sound world. How is the self/non-self distinction drawn? To consider this, the nature of our basic concept of ourselves needs to be described. Strawson claims that the fundamental aspect of this concept is that there is a single thing to which we attribute both physical features and states of consciousness (or psychological properties more generally).

Thus I, a single thing, am six feet tall *and* in pain. The question then is transformed into two; why do we ascribe states of consciousness to anything, and why to the same thing as physical states? Strawson then argues that we do not answer either question by noting the causal importance of our bodies to the character of our psychological states. This is simply not the right kind of fact to answer the question. Two other accounts of our thought about ourselves are introduced, and argued to give incoherent accounts. The first claims that contrary to appearances we do not ascribe mental states to anything – the so-called "no-ownership theory." But such an account must explain what is going on when we seem to self-ascribe states of consciousness. It proposes that we are noting the facts of causal dependence cited earlier. But, as Strawson points out, the causal dependence is of my own experiences on this body, not of all experiences on this body, and so the disavowed ascription of experiences to myself reappears.

The other account, Cartesian dualism, denies that we ascribe the two sorts of properties to the same thing; rather we ascribe the physical sort to our body and the mental sort to ourselves, a non-physical ego. The problem, according to Strawson, with this is that self-ascription of mental states presupposes the ability to ascribe such states to others, since it is in the nature of predicates to have a general application, and within the Cartesian framework there is no way to pick out other subjects to make such ascriptions. The problem is that to pick out another subject I must do so via the idea that the subject relates to a certain body as I do to mine. But this presupposes I can already think of myself. Strawson further argues that the very notion of a non-spatial particular, such as an ego is supposed to be, lacks intelligibility, for how can we understand how it is possible that there be two such which are otherwise the same when they cannot be distinguished spatially?

Strawson proposes that we have to take the concept of a person as primitive, not illusory or decomposable into elements. He then divides the predicates that we self- and other-ascribe into P-predicates, those which are unique to persons, and M-predicates, those which we share with material bodies. It follows, he thinks, that the criteria on the basis of which we ascribe P-predicates to others must be logically adequate, that is, be such that no skeptical problems can arise in the optimal case, on pain of not having an intelligible structure of concepts at all. So the philosophical problem of other minds cannot arise. Strawson adds two things: that the existence of our predicative practice here is partly explained by the special nature of action, the fact that it mixes the bodily and the mental, and that the incoherence of the Cartesian model does not imply that we cannot imagine becoming disembodied.

Strawson's description of the way we think of ourselves, as, that is, double-sided single things, seems completely correct and very important. Two lines of criticism or debate (amongst many), however, deserve mentioning. It is unclear what epistemological implications can be drawn from the plausible idea that predicates must have an intelligible potential application to a range of things. It is also unclear that Strawson should have allowed that persons can become disembodied. (For deep criticisms along the second line see C. B. Martin 1969.)

Strawson's discussion leaves quite open what should be said about a question which became central in the philosophy of mind shortly after the publication of *Individuals*, which is: what is the relation between a person's physical states and his or her mental states? Strawson at most insists that they are states of a single thing. In *Scepticism and Naturalism* (ch. 3) he considers this question and argues that there is a causal relation, rather than one of identity. He therefore rejects materialism and, in effect, espouses a type of theory that used to be called "double aspect." His interesting argument is that there will be no way to unify the mental and the physical stories, and the point of identity judgments is unification. The second premise might, of course, be questioned.

The chapter on Persons, together with that cited above, and his discussion of the Paralogisms in *The Bounds of Sense* (for which see later), constitute a profound and unified treatment of selves.

Subjects and predicates

In Part Two of *Individuals* Strawson provides a theory of the subject/predicate distinction, a task which he regards as fundamental and to which he has, repeatedly, returned. His full theory is given more recently in *Subject and Predicate in Logic and Grammar* (1974b), and I shall briefly describe it. It has a strong resemblance to his earlier account. Strawson starts with a series of what might be called marks of the subject/predicate distinction. Thus, predicates have a number of places, whereas subject expressions do not. Predicates can be negated and genuinely compounded whereas subject expressions cannot. Subject expressions are open to quantification, whereas predicates are not, as Quine suggests. Strawson's attitude is that these marks (which may need some modification too) need to be explained and do not give the basic distinction. He proposes to explain them by linking, initially in a central case, the logico-grammatical distinction to an ontological one, namely, the distinction between particulars and universals. Roughly, universals represent ways of classifying or collecting particulars. With this goes the idea that universals form structures; thus, if an object falls under one classification it follows there are others under which it does not fall; or, if it falls under one classification it follows there are others under which it does fall. Universals come in incompatibility ranges or requirement ranges. Nothing analagous applies to particulars.

The suggestion which Strawson develops is that in a language such as English, the logical features of predicates flow from the fact that the role or function of predicates is to introduce a universal, together with the second role of indicating that the referred to item exemplifies the universal. Thus the idea that subject expressions cannot be negated but predicates can is to be explained by the fact that universals form logical structures, whereas particulars do not. The inaccessibility of predicate expressions to

quantification is to be explained by the fact that they have a dual role: of introducing a universal and of indicating exemplification, for the latter is not, as one might say, a something. Strawson then extends his account beyond the basic case. No received assessment of Strawson's highly ingenious proposal has emerged.

The bounds of sense

Seven years after *Individuals*, Strawson produced *The Bounds of Sense.* In it, he analyzes, criticizes, and develops the central ideas of Kant's *Critique of Pure Reason*. The treatment of Kant is unlike that of most commentators in that it is not marked by a hagiographical reverence towards Kant, nor does it simply repeat Kant's language by way of explaining it, nor does it aspire to the length of the *Critique*. Strawson's aim is primarily to separate, insofar as it is possible, Kant's constructive and critical theses from the transcendental idealist framework in which Kant places them, and also to separate them from the outdated science and logic of Kant's time. Strawson's main claim is that transcendental idealism is incoherent, but that there are various theses that are defensible and important, defensible either in the light of what Kant himself offers or on the basis of other arguments which Strawson constructs. What, finally, is most distinctive of Strawson's treatment is the brilliant way in which he attempts to extract and defend these metaphysical and epistemological claims.

The argument against transcendental idealism occurs in Part One and Part Four, reaching the conclusion that it is incoherent in two stages. In the first, it is argued that no interpretation of Kant's idealist claims (according to which, of course, the world of space and time is merely a form of appearance, contrasting with the realm of unknowable things in themselves) is satisfactory which treats it as saying something weaker than: real objects are supersensible and we can have no knowledge of them. The second stage reasons that any model sustaining such a claim must be incoherent. Strawson argues this in various ways, but one is to suggest that we ourselves cannot coherently fit into such a picture. If, as Kant says, we merely know how we appear, is this not a genuine truth about ourselves, hence itself not merely a matter of appearance? Typically, Strawson is concerned to dig deeper and to explain Kant's adoption of the model, and he argues that its source is a distorted response to the not-at-all incoherent, indeed, according to Strawson, central to our own thinking, idea that experiances and real objects are causally related. All three elements in Strawson's response to transcendental idealism – its interpretation, evaluation, and explanation – have been disputed. (See, for example, Allison 1983 and Walker 1978.) But Strawson's independent assessment has been the stimulus of this renewed interest.

Of much greater philosophical importance, though, is Strawson's attempt to detach the central constructive and critical theses of the *Critique* from transcendental idealism and to assess them. Kant represents the purpose of the *Critique* as explaining how the synthetic a priori is possible. In his constructive reinterpretation, Strawson replaces this by the question, What features are essential to any conception of experience that we can make intelligible to ourselves? He calls the task of answering this "the metaphysics of experience." The Kantian idea is that from the notion of a self-conscious subject, who can self-ascribe its experiences, we can derive substantial conditions that must be met by the content of experiences thus enjoyed. The first condition is that it

must include awareness of what are recognizably independent objects (the objectivity thesis). Further, these objects must be recognizably spatial (the spatiality thesis), and they must satisfy various principles of permanence and causation (the thesis of the Analogies).

Strawson argues that Kant's own thesis about permanence and causation are too strong, but that rather weaker claims about semi-permanence and the necessary applicability of causal notions can be defended. However, the most crucial and brilliant, but to some extent obscure, part of Strawson's reconstruction is his defense of the objectivity thesis. Why must experience sometimes be of, or as of, objects? The argument starts, of course, from the Kantian assumption that we are dealing with the experiences of a self-conscious subject, that is, one who can ascribe to itself the experiences. For there to be any content to such ascriptions, that is, for the classification of the status of such occurrences as experiences to have a point, there must be some understanding of the contrasting status of not being an experience. This contrast or distinction can be present only if the creature's experiences, which, after all, are what sustain its concept applications, sustain the application of non-experiential categories. But that is to require that some of the experiences must present, or be of, items of a non-experiential kind. Such items must, that is, be recognized as objects.

This, in a very compressed statement, seems to be Strawson's argument. He concludes that there cannot be a genuine problem of justifying our belief in objects, for such a problem requires a vantage point where there is self-ascription of experiences without any objective judgments, which collectively await justification. No such vantage point is available. Strawson's argument clearly has affinities with Wittgenstein's private language argument. Of course, it has received much critical examination, a particularly subtle example of which is Cassam's (1995). One point is that the argument seems to rely on the assumption that categories only have contentful application in virtue of there being cases which do not fall under them, but this assumption may be rejected. Another issue is whether a thesis like the objectivity thesis relates to how experience must seem, or to how its actual objects must be.

In subsequent chapters Strawson considers what can be defended from Kant's arguments in the Analogies. Here the strong Kantian thesis cannot be sustained, but weaker versions are, according to Strawson, defensible. Thus, for experience to be recognizably of objects, the experiencer must be able to distinguish the temporal history of the objects from those of the experiences, which requires that the objects as experienced yield a framework in terms by which to keep track of them, both spatially and temporally; and this requires that the experiences ground the application of concepts of enduring objects (in one sense, that is, of substances). Further, as Strawson puts it, objects have to be understood as the ground of "compendia of causal laws." So the application of causal notions is also required. The latter stages of argument here both resemble and develop the arguments in *Individuals*.

The *Bounds of Sense* also analyzes the critical program of the Dialectic. It does so in a deep and very illuminating way, especially in connection with the illusions of rational psychology as exposed in the Paralogisms. Strawson sees Kant's achievement in that section as refuting attempts to infer that selves are special non-physical things from what are undoubtedly special epistemological features of self-knowledge. The major incompleteness in Kant's account is his reluctance to settle for an embodied self, limited

as he is by his transcendental idealism. The result is that Strawson's book has had an enormous impact on both Kantian scholarship and recent metaphysics.

Responses to skepticism

Strawson has had a number of interests as an epistemologist. He has not attempted to provide an analysis of knowledge, but he has tried to describe its structure, and especially to give an accurate account of the role of perception within that structure. However, his influence has been greatest in his role as opponent of skepticism. But his attitude to skepticism has evolved, and I want to describe briefly four main stages.

In *Introduction to Logical Theory*, he argued in relation to the particular case of induction, that when the skeptic claims that no justification has been provided, there is no intelligible and possible thing that can be understood by "justification." The correct response to skepticism is, therefore, not to try to provide a justification but to see that there cannot be such a thing. Why cannot there be such a thing? The answer to this emerges when one of a range of fuller specifications of the task is given. Thus, one way of understanding the notion of justification here takes it to require showing that inductive support is really deductive. Clearly this is absurd, in that it requires an obliteration of the very distinctive method of support that raises the problem in the first place. Another understanding is that a justification would amount to showing induction is a reasonable procedure; but this is not something that needs to be *shown*, since "being reasonable" precisely means following inducation. Another understanding is that a justification would be a proof that induction is bound to work; but such a proof is impossible. There is, according to Strawson, no coherent demand here. The structure of Strawson's argument here leaves room for someone to find a coherent interpretation of the skeptic's claim, and subsequent discussion has either aimed at doing that, or at disputing the analytic claims Strawson himself makes.

In *Individuals*, Strawson seems to have argued that skepticism is incoherent in that possession of the concepts that the skeptic needs to identify the topic of his own skepticism itself requires a non-skeptical attitude towards the (best) bases of application for those concepts. For example, having mental concepts involves understanding their application across a range including others and this requires that the criteria of application to others are logically adequate. Strawson argues in a similar way in respect of across-time identity judgments. The much-debated question about such claims is whether it is shown that the adoption of a skeptical attitude is really inconsistent with understanding.

The third sort of response that Strawson has explored, in *The Bounds of Sense*, are transcendental arguments. The idea is that the skeptic himself supposes that some conceptual applications are possible, for example, a non-committal description of experience. But it is argued that the very ascriptions that the skeptic is prepared to make presuppose the application of the concepts that he is skeptical about. Thus, Strawson argues, as we have seen, that the self-ascription of experiences requires judgments about objects. Now, there is no a priori reason to hold such an argument could not be correct, but any transcendental argument of more than minimal length runs the risk of overlooking a way in which the skeptic's concepts can have application without needing the ones in dispute.

Finally, in *Scepticism and Naturalism*, Strawson suggested another response. Arguing as he thinks in the spirit of Hume but especially of Wittgenstein, he draws a distinction between real doubts which are worth engaging with and unreal doubts which cannot assail anyone and which are not worth responding to. The traditional skeptical doubts fall into the second category. No one ever seriously wondered whether there is an external world. Strawson suggests that we do not need therefore to *argue* against skepticism. This very bold and historically informed response has not been popular, the ground being that it is not obvious that the non-persuasiveness of an argument means it need not be engaged with. It might also be wondered whether the historical roots for such a response are not more Lockean than Humean (see MOORE).

Still, in postwar analytical philosophy no agreed response to skepticism has emerged, and Strawson can be credited with the development of a number of the candidates still being investigated.

Freedom and resentment

Strawson has written little about moral philosophy, his attitude to it being, perhaps, somewhat similar to that expressed by C. D. Broad, when he reputedly said that the whole of moral philosophy could be written on the back of a postage stamp. However, one of Strawson's essays on moral philosophy, namely, "Freedom and Resentment" (in 1974a, but first published in 1962) has been extremely influential. Interestingly, Strawson's argument in this article bears a close relation to the case against skepticism in his 1985 book.

Strawson's aim is to find a position in the debate about determinism and responsibility that avoids incompatibilism (which Strawson relabels "pessimism") but without distorting the nature of our moral view of ourselves and others in the way in which, according to Strawson, standard compatibilists do. Standard compatibilists observe that our practices of punishing and praising would have a utility even if determinism was accepted. Strawson claims that viewing the issue this way over-intellectualizes the basis for such practices, which is not reflectively shaped by considerations of utility, but rather is the upshot of certain central reactive attitudes engendered in the course of ordinary human life. Examples are resentment and gratitude, which are directed at others, but also guilt and remorse, which are self-directed. The reconciling project is based on the following claims: such attitudes and feelings are produced in us in the course of our normal participation in human life, with its necessary engagement with others; the attitudes can be suspended in exceptional circumstances, for example, when dealing with people who are palpably mentally abnormal. They cannot, however, be universally suspended, because they are integral to human relationships that we cannot abandon. And when they are suspended in the limited cases where this is possible it is not because we see the cases in the light of a general conviction in determinism, but rather because of more specific reasons, which vary from case to case. Strawson draws from the alleged impossibility of abandoning such reactions, and the absence of a dependence of them on a rejection of determinism, his central compatibilist conclusion that determinism is no threat to their legitimacy. He draws from the claim that they are natural and not governed by consequentialist considerations the conclusion that standard compatibilists have distorted the character of our moral responses.

Strawson's paper was important because it represented a novel compatibilist approach. The central issue it raises, though, resembles that raised by his observation that people are not persuaded by skeptical arguments. Does the fact that people will carry on believing or doing something show that it is beyond criticism, or legitimate for them to do so?

Conclusion

The most striking aspect of Strawson's philosophical career has been his extraordinary fertility, combined with the consistent depth and clarity of what he has produced. In the present essay, there has not been space to survey many aspects of his work. A final indication of his importance, though, is his influence on the very best philosophers of the generations after his, of whom I wish to mention only two. The first is Gareth Evans, whose book *The Varieties of Reference* is clearly colossally influenced by Strawson. The second is John McDowell, who in the preface to *Mind and World* pays eloquent tribute to Strawson. What might be called a Strawsonian tradition has emerged.

Bibliography

Works by Strawson

1952: *Introduction to Logical Theory*, London: Methuen.
1959: *Individuals*, London: Methuen.
1966: *The Bounds of Sense*, London: Methuen.
1971: *Logico-Linguistic Papers*, London: Methuen.
1974a: *Freedom and Resentment and Other Essays*, London: Methuen.
1974b: *Subject and Predicate in Logic and Grammar*, London: Methuen.
1985: *Scepticism and Naturalism: Some Varieties*, London: Methuen.
1992: *Analysis and Metaphysics*, Oxford: Oxford University Press.
1997: *Entity and Identity*, Oxford: Oxford University Press.

Works by other authors

Allison, H. E. (1983) *Kant's Transcendental Idealism*, New Haven, CT: Yale University Press.
Cassam, Q. (1995) "Transcendental Self-Consciousness," in *The Philosophy of P. F. Strawson*, ed. P. K. Sen and R. R. Verma, New Delhi: Indian Council of Philosophical Research.
Hahn, L. E. (ed.) (1998) *The Philosophy of P. F. Strawson*, La Salle, IL: Open Court. (Includes an intellectual autobiography by Strawson, a collection of essays discussing his work, and his replies.)
Martin, C. B. (1969) "People," in *Contemporary Philosophy in Australia*, ed. R. Brown and C. D. Rollins, London: Allen & Unwin.
Van Straaten, Z. (ed.) (1980) *Philosophical Subjects*, Oxford: Clarendon Press. (A distinguished collection of papers on Strawson with his replies, the papers including G. Evans's "Things Without the Mind," his profound response to ch. 2 of Strawson's *Individuals*, and J. McDowell's "Meaning, Communication and Knowledge," a response to Strawson's lecture on meaning and truth.)
Walker, R. (1978) *Kant*, London: Routledge and Kegan Paul.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

27

Philippa Foot (1920–)

GAVIN LAWRENCE

Philippa Foot is among the handful of the twentieth century's very best moral philosophers. Her achievement consists not so much of truths presented as of her distinctive voice in philosophy. In this way, she is like Moore or Rawls, or most pertinently Wittgenstein. To read her is immediately to struggle with the real stuff of the subject, to the highest standards; the subject is not the same for one again.

Her work divides into several, diversely overlapping, strands: the major themes of ethics, such as its objectivity and its rationality; middle range issues, such as freedom of the will, virtues and vices, the critique of utilitarianism, and moral dilemmas; more specific ethical distinctions and problems, such as the doctrine of double effect, abortion, euthanasia, and capital punishment. I will focus on the major themes.

Her treatment of the issues of morality's objectivity and of rationality falls into three phases. These phases relate to three Humean (or neo-Humean) orthodoxies: (1) the fact/value distinction, (2) the practicality of morality, and (3) the end-relative conception of practical reason. Roughly, Foot starts by rejecting (1) while accepting (2) and (3). She then rejects (2) as well. Finally she rejects (3) in favor of a more Aristotelian conception of practical reason and comes to reassert (2).

The first phase (1950s to mid-1960s): the Wittgensteinian defence of the possibility of naturalism

From the first, Foot has taken mainstream contemporary moral philosophy to be dominated by two of the three Humean propositions:

(1) *The fact/value distinction* (anti-naturalism) assumes, against the naturalist, that there is some logical gap between fact and evaluation – between "is" and "ought." Evaluations go beyond the natural facts. And, in the subjectivist version, they require a contribution from the subject. If so, evaluative judgments, unlike factual ones, are not wholly responsible to the world, and evaluative argument may *break down* in a way that factual argument cannot: two opponents may agree about all the facts, and yet commit their will differently and so be left in a bare opposition of will or attitude, without any rational error (see ANSCOMBE).

Three versions of (1), "the breakdown theory," particularly concern us. (1a) *Radical subjectivism* claims that no particular, conceptually restricted, range of facts is either

necessary or sufficient evidence for a certain evaluative predicate (i.e. no content restriction). (1b) *Restricted subjectivism* claims that, while necessary, such a range of facts is never sufficient. (1c) *Partial subjectivism* claims that, while necessary and sometimes sufficient, such a range of facts is not *always* sufficient.

(2) *The practicality of morality* is the orthodoxy that morality is somehow "practical" or action-guiding. Although (1) secures a division between fact and value, it allows one to explain the second orthodoxy. It is because morality is a matter of value, not of fact, that it can be action-guiding. It may be that the practicality of morality may be a matter either of its motivational efficacy, or of its rationality and thus the claim that moral considerations universally motivate, or that they universally constitute reasons. (The latter raises the further issue of the nature of practical reason, and thus the relevance of (3).) Further differences exist over the modality of the thesis: does morality just happen to be universally motivating/reason-giving, or is it necessarily, or essentially, so? and over whether the motivation/reason yielded by morality is supposed overriding (or authoritative).

In her first phase – most notably in "Moral Arguments," "Moral Beliefs," and "Hume on Moral Judgement" (all in 1978) – Foot argues both that (1) has not been made out by its anti-naturalist proponents, and that (2) doesn't in fact require it.

Her initial target is (1a) radical subjectivism. According to this view, there are supposedly *no* content restrictions on what can be held to be morally good, or a moral principle, or a moral code. Consistency apart, a person is free to discount the facts or grounds anyone else takes as evidence for something's being good, and is free to count as evidence facts that no one else acknowledges as evidence. This personal freedom to decide relevant grounds – what Foot terms "the private enterprise theory" of morality – seemingly risks making evaluative predicates meaningless. Isn't a predicate that can be freely pasted anywhere necessarily uninformative? What initially saves the subjectivist is the "linguistic turn," so influential in the 1940s and 1950s. This is the recognition that language is multi-functional, and that, in particular, it has other purposes besides the descriptive or informational. Thus, the anti-naturalist need not follow Moore in holding that "good" is descriptive of a non-natural property, rather than of a natural one. Rather the anti-naturalist holds that the primary use of "good" is not to describe the world, but to express an attitude or to recommend.

Foot argues that radical subjectivists do not prove their case. She begins by considering a middle level, or thick, predicate, such as "rude," that her opponents would likely concede is evaluative (i.e. to have an expressive or action-guiding function). But, as she points out, if one uses the concept of rudeness, one isn't free to take just anything one likes as evidence for rudeness (e.g. walking slowly up to a door), or to reject just anything either (e.g. being spat on), any more than one is free to decide what is and isn't evidence for a brain tumor. (Of course, given a suitably special background story, behavior that normally is not rude may be rude, for example pointedly walking slowly when asked politely to hurry, and vice versa.) Foot then claims that, for all the opposition has argued, what holds for "rude" may hold for *all* evaluative concepts, including the more abstract "thin" ones. For all they argue, there may yet be the "very tightest of relations" between fact and value. Evaluative concepts are, like any others, criteria-governed concepts: they have "definitional criteria" which lay down what is and isn't relevant evidence for them. Leave these criteria behind and you leave behind the concept. (In

"Moral Beliefs" she puts the point in terms of an object being internally related to the attitude it is an object of. One cannot feel proud of just anything: one has, say, to be thinking of it as an achievement of one's own.)

At this point subjectivists can go in one of two directions to avoid Foot's position.

First, they can accept Foot's point for "thick" evaluative terms, but reject it for "thin" ones. They may claim that we are free to decide what counts as benefit or harm; or they may admit that there are criteria, or rules of evidence, for our existing moral code and our existing moral terms, but claim we are always free to invent new moral terms, new virtues, or moral codes.

To this, Foot replies in a manner reminiscent of Wittgenstein's *Philosophical Investigations* §261, that factual constraints apply here too. Not just anything can count as a benefit or a harm, nor just anything count as a virtue or *moral* code. The claim that no one should look at hedgehogs in the light of the moon could not count as a *moral* principle without a special background. More generally, even the word "moral" is content-restricted. Not just any alien, or trivial, code can count as a moral one. Roughly, moral considerations must relate to human good and harm. (Foot does not claim to have thoroughly elucidated the definitional criteria at work in our moral predicates.)

Alternatively, subjectivists can opt for a *restricted subjectivism*, by using proposition (1b). According to this view, Foot's conceptually restricted descriptive conditions (a) are necessary for an application of an evaluative predicate, and (b) sufficient for a merely descriptive (or "inverted comma") application of them. But, it claims, these conditions do not suffice for a properly evaluative use. To think they do is to miss the very point of evaluation, namely, that therein agents contribute something of their own after the facts are settled, be it a commitment of will or intention, feeling, or attitude, etc. It is this further element, of its nature linked to a tendency to act, that is needed to secure morality's practicality, as in (2). It is something entirely up to agents, and cannot be logically required of them by the world.

In "Moral Beliefs," part II, Foot argues that this position too is mistaken: it puts "the practical implication of value words in the wrong place." Her argument, as I read it, contains a carrot and stick. The carrot is the offer of a more plausible account of morality's practicality. Injury, she suggests, offers a helpful parallel. Once we agree that not just anything can be called an injury, we can see that the *reason* for us to avoid injury is not that "reason-givingness" is built into the evaluative use of "injury." Rather, it is simply that certain kinds of things count as injuries. I have a reason not to poke a sharp object in my eye not because I find myself prepared to call this an "injury" in a full evaluative or full action-guiding, sense, but because I won't be able to see, and, as things are, I need to see. Similarly the connection between moral judgment and the will does not lie in the will's commitment being a condition constitutive of evaluative use, but in the content of moral judgment. It is the facts about what the virtues are, given the conditions of human life, that secure that there is reason for each of us to be virtuous, and act virtuously.

The stick is to query how any such extra element – be it attitude, disposition to choose, self-addressed imperative, or whatever – could possibly perform this role of rationalizing actions or character traits. For all suggested candidates seem obviously

unnecessary, as far as reason-givingness goes. I may know that courage is a virtue and that I have reason to be courageous, but, coward that I am, have no commitment, or whatever, to mending my ways.

The alternative account of morality's practicality, (2), interprets it as a matter of universal rationality, not of universal motivation. (Foot supposes this the more plausible version.) And, by its very nature, it ushers back on stage another, much older, opponent, the immoralist, who, like Plato's Thrasymachus agrees that morality's rationality is settled by facts about the virtues, but queries whether these favor the recognized virtues, and justice in particular; for justice, on the face of it, is "another's good and self harm." The rest of "Moral Beliefs" attempts to answer this opponent.

But which facts about the virtues and virtuous action would show that there is reason to pursue them? The specific form of Foot's alternative account, and of her reply to the immoralist, is controlled by her assumption of an *end-relative* conception of practical reason, namely the third orthodoxy.

(3) A consideration C is a reason for agents if, and only if, it serves something they desire or care about, that is, it is a reason only in relation to their ends. Given this view of what makes something a reason, we all have reason to be virtuous and to act virtuously if and only if the facts about the virtues and virtuous actions show them to connect up with what each person happens to want or care about. This connection could be instrumental or constitutive. Moral considerations may be reasons either because they further an agent's non-moral end, or because they are constitutive of achieving some moral end of the agent's. Foot supposes that not everyone has moral ends, and that the only end universally shared is a non-moral one of self-interest (albeit not necessarily selfish). Thus she feels that, to defend the universal rationality of morality (her commitment to (2)), she has to demonstrate a kind of moral instrumentalism: that, as things are, the virtues further self-interest.

The resulting position has its problems. One that immediately occupies Foot is the defence of justice. It is not difficult to make a general case that justice furthers an agent's self-interest; but what of the "tight corner," the particular case where to act justly an agent has to lay down her life? How can justice here be more to the agent's advantage or self-interest? The solution she offers was a version of Hume's but, by her own account, it was in part dissatisfaction with this that leads her next to abandon (2). It is not, however, until the third phase that Foot locates the real culprit, the neo-Humean view of practical reason. Only then is the restricted subjectivist presented with a proper alternative account, and the immoralist with an adequate response.

The second phase (1970s): unease over morality and the rejection of (2)

The mark of Foot's second phase is the suspicion that our ordinary moral thought and language contains elements of fiction, in its assumptions of complete objectivity and of rationality or authority. In "Morality and Art" (1970) Foot still rejects the fact/value distinction, (1). The definitional criteria built into the concept of the moral constrain what can count as a moral code or as morally good and explains why so many moral

judgments can be proved from the facts (e.g. that Hitler's treatment of the Jews was wicked). The core of morality is objective, and its truth non-relative. Nonetheless these criteria are not so stringent as to rule out subjectivism entirely. For example, regarding abortion and euthanasia, Foot suggests that different people could, subjectively, choose to go by different principles, and each choice would equally count as moral. And at such points there could be that very kind of breakdown in moral argument that Foot earlier denied. Foot thus embraces *partial subjectivism*, (1c).

An analogous possibility is presented for relativism, where different elective moral principles may be peculiar to different groups. At these points moral truth would be objective but relative to the standards adopted by a particular community.

In the 1978 Postscript to "Morality and Art" and in "Moral Relativism" (1979), Foot is less confident about these points and holds that they cannot be settled without a firmer grip on the nature of the definitional criteria and of certain key concepts such as having a value and happiness. This is still unfinished business (see the end of "Does Moral Subjectivism Rest on a Mistake?" (1995)).

More scandalously still, in "Morality and Art" and then in "Morality as a System of Hypothetical Imperatives," (both 1978) Foot challenges the orthodoxy that everyone *should* be moral and act morally. She distinguishes two uses of "should." One is a non-hypothetical, or desire-independent, use: it says what is required by a certain point of view or system. And in this sense it is tautological that one should, morally speaking, be moral. The other use of "should" is reason-giving. Clearly this is the use at issue here. But, Foot argues, the claim that, whatever their desires and interests, everyone has reason to be moral and act morally lacks a sense; for it implies that moral considerations have a magically automatic reason-giving force. Instead we should concede that moral, like other, considerations offer reasons only hypothetically, that is, on the condition that the agent happens to have the appropriate ends. (Note that now she supposes moral considerations are properly reasons only for those who have moral ends, that is, via a constitutive, not an instrumental, connection.) If so, people who lack moral ends will have no reason to act morally, and to say that they should so act, or should have such ends, is mere bluff. And so Foot abandons the claim of morality's universal rationality (2), viewing it as another piece of moral fiction, complete with a fictitious linguistic use.

Morality then turns out to be inescapable in one sense but not in another. The application of the moral predicates – just, courageous, mean, cruel, etc. – is an objective matter. But whether moral considerations are reasons turns on the subjective matter of what the agent happens to care about. In short Foot's position is that of moral objectivity, rational subjectivity. Once again the controlling assumption is her commitment to the end-relative conception (3), as being the only non-mysterious view of reason.

The third phase (1980s–1990s): rejecting (3); objective morality, objectivity rationality, and the facts of human life

Foot's more recent phase is most apparent in "Rationality and Virtue" (1994) and "Does Moral Subjectivism Rest on a Mistake?" (1995). Central to it is the replacement of the subjective theory of practical reason, (3), by an objective one. This allows Foot

both to reaffirm morality's rationality, (2), by offering the restricted subjectivist an alternative account of it that is more convincing than her earlier "instrumentalist" defence; and also to use the objectivity of reasons to get at what is really wrong with immoralism and thus successfully to conclude her long struggle with Nietzsche.

Foot now claims it is a mistake of strategy to start from some preconceived theory of practical reason, such as the maximization of perceived self-interest, or desire-satisfaction (cf. (3)), and then try to show that moral action is rational in its terms. Instead the rationality of moral action is on a par with that of self-interested action: they are not rival theories, but different parts, of practical rationality.

There are three main elements in her new position. First, she elucidates the concept of a moral virtue as an excellence that ensures that an agent is good in respect of action (and feeling). A virtue does this by being a disposition correctly to count certain considerations as reasons, and then to act on them, that is, to do well in respect of acting on reasons. So a moral virtue is goodness in reason-recognition and reason-following. As such it is part of what it is for practical rationality to be in good order. Given this formal connection between moral virtue and rationality, there can be no question of whether a virtuous act is rational. (If justice is a virtue, then an unjust action will thus be contrary to practical reason.)

However the connection is *only* formal. We are still left to determine what actually are the moral excellences of acting on reasons (e.g. that justice is one). Foot then argues that this is settled, quite objectively, by facts about human nature and life. We readily grant that there are objective factual evaluations of what count as excellences and defects in such faculties as sight or memory – be it in an elephant, owl, or human – on the basis solely of the natures, needs, and forms of life of the respective species (e.g. lack of good day sight is not a defect in an owl). Foot calls this "autonomous species-dependent goodness." It applies equally to behavioral operations: nest-building, hunting in packs, etc. And, allowing for certain differences, Foot argues, the same basis of evaluation applies to the human operation of acting on reasons, to determine its excellences and defects. Because of what we are and what we do, we need such things as being able to bind others by promises, and mutual helpfulness. So we need to recognize and follow the reasons they present. These general facts of human nature and form of life fix what considerations are reasons for humans, and do so quite objectively, that is, regardless of whether or not some individual (e.g. an immoralist) recognizes them.

Finally, Foot extends the same account to the other part of practical rationality, prudence. That considerations of self-interest are reasons is validated once again by general facts of human nature and life: that adult humans plan and look out for themselves better than others can for them. The basic ground of the rationality both of moral virtues and of prudence is the same, allowing us to have in this respect a unified theory.

The general shape is neo-Aristotelian, albeit with distinctive elements. Much is controversial: the treatment of self-interest, the normative view of human nature, the swiftness of the answer to immoralists (who will complain they see no reason to be "good humans"). More needs saying, and Foot's book (forthcoming) will say more. It will, I believe, offer the twenty-first century a much better start than Moore's *Principia Ethica*, of 1903, did to the twentieth.

Bibliography

Works by Foot

1970: "Morality and Art," *Proceedings of the British Academy* 56, pp. 131–44. (Reprinted with postscript in *Philosophy As It Is*, ed. M. Burnyeat and T. Honderich, Harmondsworth: Penguin Books.)
1978: *Virtues and Vices*, Oxford: Blackwell Publishers.
1979: *Moral Relativism*, Lindley Lecture, Kansas: University of Kansas Press.
1994: "Rationality and Virtue," in *Norms, Value, and Society*, Vienna Circle Institute Yearbook, Amsterdam: Kluwer.
1995: "Does Moral Subjectivism Rest on a Mistake?," *Oxford Journal of Legal Studies* 15, pp. 1–14.
forthcoming: *The Grammar of Goodness*, Oxford: Clarendon Press.

Work by other authors

Hursthouse, R., Lawrence, G., and Quinn, W. (eds.) (1995) *Virtues and Reasons: Philippa Foot and Moral Theory*, Oxford: Clarendon Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

28

Ruth Barcan Marcus (1921–)

M. J. CRESSWELL

After undergraduate work in mathematical logic with J. C. C. McKinsey, Ruth Barcan proceeded to Yale to work with F. B. Fitch and to produce a Ph.D. dissertation in modal logic. In the mid-1960s she established the philosophy department at the University of Illinois at Chicago. Later she moved to Northwestern University and to Yale University. She has been President of the Western Division of the American Philosophical Association and of the Association for Symbolic Logic and has been the recipient of many grants and awards, including the Medal of the Collège de France and an honorary doctorate from the University of Illinois.

Ruth Marcus is the author of the earliest published work in modal predicate logic, that is, modal propositional logic extended with quantifiers and predicates. Her first article appears in volume 11 of *Journal of Symbolic Logic* (Barcan 1946). (Rudolf Carnap's "Modality and Quantification" appears later in the same volume.) It is perhaps surprising that although modal logic in the form in which we know it today dates from an article in *Mind* by C. I. Lewis in 1912, it was not until 1946 that there is any consideration in print of what happens when quantifiers and predicates are added. In her first article Marcus considers the predicate extensions of the systems of modal logic known as S2 and S4. S2 was Lewis's preferred modal system, and is the one developed in detail in Lewis and Langford 1932. Many of the theorems of modal predicate logic are simply instances of non-modal predicate logic or of modal propositional logic, but some are not. Of these an axiom that Marcus adopted was $\Diamond\exists xA \Rightarrow \exists x\Diamond A$ where $\Diamond$ is the modal possibility sign, and $\Rightarrow$ is the sign for "strict implication," and where $A \Rightarrow B$ is defined as $\sim\Diamond(A \,\&\, \sim B)$. Arthur Prior (1956: 60) called "$\Diamond\exists xA \Rightarrow \exists x\Diamond A$" the "Barcan Formula" and the name has stuck. Whether or not to include the Barcan Formula as a truth of modal logic has been a matter of much controversy. Prior (1957: 26) argued that the temporal version of this formula should be rejected on the ground that a sentence like "it will be that someone will be alive in 2150" does not entail that there is anyone who will be alive then. (Prior's example was flying to the moon, but that is now a little anachronistic!) Marcus herself has entered this discussion by defending the formula. She protested (in Marcus 1962) against the reading of $\exists xA$ as "there exists an x such that A," and proposed instead (p. 252) that it be read as either "some substitution instance of A is true" or, alternatively, as "there is at least one value for x for which A is true." Such a reading allows the quantifier to speak of things which, in the

temporal case, no longer exist or do not yet exist, and in the modal case do not exist but might have.

The alternatives are slightly different. The second can be treated in possible-worlds semantics if the domain of the quantifiers is expanded to include possibilia (things which exist in other worlds but may not exist in ours). The first alternative is the one Marcus herself has championed; it is to adopt a substitutional interpretation of quantification, whereby $\exists xFx$ is true iff Fa is true for some constant a. Marcus's views on this have been developed in more recent work, most of which has been reprinted in *Modalities* (1993). On the substitutional interpretation the Barcan Formula is uncontroversially true, since if $\Diamond\exists xFx$ is true then $\exists xFx$ might have been true. But then, by the substitutional account of the quantifier, some instance Fa of Fx might have been true, and so $\Diamond Fa$ will be true, and so $\exists x\Diamond Fx$ will be true. Nothing is said here about whether or not a exists in this or any other possible world (or perhaps more accurately whether the name "a" refers to anything) though Marcus is not unsympathetic to a defense of the formula in terms of a fixed domain, and undertakes such a defense herself, in *Modalities* (1993: 21f.). There is of course a question of what it means to say that Fa is true. Marcus's attitude is that this is an issue about the interpretation of singular statements involving names, and not an issue about quantification, and that it is a virtue of the substitutional interpretation that it divorces these two questions. It is no longer possible to follow Quine and locate the commitment in the quantifiers via the satisfaction of an open sentence of the form Fx.

In one of Marcus's early papers (Barcan 1947) there is a theorem that if x and y are identical then this is necessarily so. Marcus derives this from the standard principle in ordinary non-modal predicate logic with identity that if $x = y$ then any two formulae that differ only in that one has free x in some places where the other has free y are equivalent. Yet it seems that although nine and the number of the planets are identical, for there are nine planets, this identity is not necessary, for there might have been more or fewer. It is now a commonplace that this puzzle is easily solved by Russell's theory of descriptions without giving up Marcus's theorem about the necessity of identity, as was shown by Arthur Smullyan in the 1948 *Journal of Symbolic Logic*. Yet in 1962, when Marcus presented a paper on the role of identity in modal languages, it was (as she says in her introduction to the paper in 1993: 3) a mistaken assumption on her part that Smullyan's paper was fully appreciated. The appendix to this paper, based on a taped discussion between Marcus, Quine, Kripke, Follesdal, and others, makes it clear how difficult it was to come to grips with these issues in the days before the power of possible-worlds semantics for modal logic was widely understood. Some of Quine's worries about quantified modal logic are on the grounds that it leads to "Aristotelian essentialism." Although noting that modal predicate logic is not committed to essentialism Marcus concedes that it is compatible with it, and defends a form of essentialism in which the modalities are understood causally (1993: 67–70).

Marcus is a strong supporter of the causal theory of names. Proper names, on this view, have as their meaning nothing more than the object they denote, and they are able to have this meaning in virtue of a causal connection between an initial "dubbing" of the object and subsequent uses of the name. (It is somewhat unfortunate that a dispute has grown up, to which neither Marcus herself nor Kripke is a party, about the historical priority between her and Saul Kripke on the treatment of names as mere

"tags" rather than as descriptions. Whatever might be said about that issue does not in the least detract from the value of Marcus's views on these matters.) If quantification depends on naming, and if naming demands a causal connection, then it is difficult to see how quantification could ever apply to things that do not exist. The causal view of course goes very happily with the view that ontology is linked to reference. "Actual objects are there to be referred to. Possibilia are not" (1993: 205).

Many attempts to defend possibilia appear to be based on the view that we can refer to non-actual objects like Pegasus and Sherlock Holmes. Marcus rightly rejects such attempts. She is also rightly points out (1993: 194) that sentences such as "the winged horse does not exist," when analyzed according to Russell's theory, involve no reference to a possible but non-actual winged horse (see RUSSELL). But she is well aware that the difficult cases are not these. They are sentences like "There might have been more things than there are," where there is no question of referring to any of them. A philosopher with a more realistic attitude to possible worlds than Marcus might no doubt say that any possible but non-actual object can be referred to, but of course only in a world in which it exists, not in our world. Marcus's attitude to such views is not sympathetic and she speaks of such semantics for quantified modal logic as providing it with "a different subject matter from that of non-modal logic" and as not being "a straightforward extension of standard predicate logic" (1993: 191).

A third theme in Marcus's work is connected with contradictions. She defends the claim that moral dilemmas are real, but need not threaten the consistency of a moral code. If we think of a moral code as a set of sentences, then a code will be inconsistent if and only if all its members cannot be simultaneously true. In this sense a consistent code may well allow the possibility of dilemmas. For suppose that the code says that if p then Oq ("Oq" means that q is obligatory) and that if p then O~q (it is obligatory that not q). If this is formalized as $(p \rightarrow Oq) \& (p \rightarrow O\sim q)$ then even if we grant that Oq and O~q are jointly inconsistent it still does not follow that $(p \rightarrow Oq) \& (p \rightarrow O\sim q)$ is inconsistent since $(p \rightarrow Oq) \& (p \rightarrow O\sim q)$ will be true if p is false. In describing a game, Marcus says "a game might be so complex that the likelihood of its being dilemmatic under any circumstances is very small and may not even be known to the players" (1993: 134). If I have understood her correctly, her point is that although it is possible that the moral life might land us in situations where we cannot do the right thing, yet moral dilemmas may often be avoided provided that the world cooperates. She endorses a second-order moral principle (pp. 139f.) to the effect that we should so order our lives that as far as possible they are in fact avoided.

Her attitude to believing the impossible is different. An impossible proposition is a proposition true in no possible world. So there is no possible world which would be the way things are if a belief in an impossibility were true. Yet it would seem that we often do believe contradictions. Marcus discusses Kripke's well-known example of Pierre who believes that London (the city he knows as "Londres") is pretty, and also that London (a city he has come to know as "London") is not pretty. Since this conjunction is a contradiction Marcus claims that Pierre cannot have this belief (1993: 158).

Marcus's work (excluding the early technical articles) has been collected in *Modalities* (1993) and essays in her honor appear in *Modality, Morality and Belief* (Sinnott-Armstrong et al. 1995). However, her influence is not restricted to her writings, and perhaps does not even primarily come from her writings. At the institutions

at which she has taught she has influenced several generations of students who have become leading philosophers; and her role in the international philosophical community has been no less significant.

Bibliography

Marcus's earliest work is listed under her maiden name, Barcan. Of her later articles, those that are referred to specifically are listed here; others can be found in the 1993 collection.

Works by Marcus

Barcan, R. C. (1946) "A Functional Calculus of First Order Based on Strict Implication," *Journal of Symbolic Logic* 11, pp. 1–16.

——(1947) "The Identity of Individuals in a Strict Functional Calculus of Second Order," *Journal of Symbolic Logic* 12, pp. 12–15.

Marcus, R. B. (1962) "Interpreting Quantification," *Inquiry* 5, pp. 252–9.

——(1993) *Modalities*, New York: Oxford University Press.

Works by other authors

Lewis, C. I. (1912) "Implication and the Algebra of Logic," *Mind*, new series 21, pp. 522–31.

Lewis, C. I. and Langford, C. H. (1932) *Symbolic Logic*, New York: Dover Publications.

Prior, A. N. (1956) "Modality and Quantification in S5," *Journal of Symbolic Logic* 21, pp. 60–2.

——(1957) *Time and Modality*, Oxford: Clarendon Press.

Sinnott-Armstrong, W. (1995) *Modality, Morality and Belief, Essays in Honor of Ruth Barcan Marcus*, ed. D. Raffman and N. Asher, Cambridge: Cambridge University Press. (Contains a full bibliography of Marcus's work.)

Smullyan, A. F. (1948) "Modality and Description," *Journal of Symbolic Logic* 13, pp. 31–7.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

29

John Rawls (1921–)

NORMAN DANIELS

John Bordley Rawls, who developed a contractarian defense of liberalism that dominated political philosophy during the last three decades of the twentieth century, was born in Baltimore, Maryland. In 1939, he left his home town to attend Princeton. He served in the Pacific (1943–5), and returned to Princeton to receive his Ph.D. ("A Study in the Grounds of Ethical Knowledge") in 1950. He taught briefly at Princeton, Cornell, MIT, and then, for thirty years, at Harvard. He married Margaret Warfield Fox, a painter, in 1949. They raised two sons and two daughters and have lived for many years in Lexington, Massachusetts.

Rawls's enormous influence in philosophy, law, economics, and political science is largely traceable to his major work, *A Theory of Justice* (1971). According to one survey, it is one of the five most cited philosophical books of the twentieth century. In contrast to the dominant emphasis in twentieth-century ethics on the analysis of moral language and on topics in metaethics, *Theory* argued rigorously for substantive moral principles and discussed their implications for the design of basic social institutions. (See, e.g. ANSCOMBE, AYER, FOOT, HARE, MOORE, STEVENSON; cf. CHOMSKY and POPPER.) This normative stand encouraged other work on justice, as well as on other areas of applied ethics, and it explains the relevance of Rawls's work beyond philosophy. Rawls's influence is also the result of his dedication to teaching. He has trained many of the leading philosophers in ethics and political philosophy over several generations.

Rawls's first publication, in 1951, "Outline of a Decision Procedure for Ethics," though deriving from his Ph.D. thesis, expresses a lifelong theme in his work. It proposes a procedure for selecting and justifying ethical beliefs and principles from among the diverse views people hold. As he struggled with this problem over the next two decades, he narrowed the scope of his proposed solution. In *A Theory of Justice*, Rawls focuses only on a procedure for selecting among competing principles of justice, not moral principles quite generally.

In *Theory*, Rawls uses a hypothetical social contract (the Original Position) to argue for principles of justice different from the utilitarianism that has long dominated Anglo-American philosophy. Deliberating behind a "veil of ignorance" that blinds them to distinguishing and potentially-biasing facts about themselves, rational contractors choose principles that protect certain basic liberties, including the effective exercise of political liberties, guarantee fair equality of opportunity, and permit inequalities (measured

by an index of primary social goods) only if the inequalities work to make those who are worst off as well off as possible (the Difference Principle). Together these principles regulate the basic structure of society and produce a form of egalitarianism that Rawls calls "democratic equality." Because these principles are chosen in a situation that is fair to all contractors, Rawls labels his view "justice as fairness," by which he means procedural fairness.

In addition to being the rational choice of contractors, the principles must also meet two other conditions. They must match or cohere with "our" considered judgments about justice in (wide) "reflective equilibrium." They must also be feasible in the sense that people raised in a society governed by them would find the system to be more stable, with less strain of commitment, than alternatives.

Following the publication of *Theory*, there was extensive critical response in philosophy journals and books, as well as in related fields. Rawls engaged actively with this critical literature over the next two decades. In the first few years after *Theory*, he defended his focus on the basic structure of society, his coherentist account of justification, his use of the primary social goods, his argument for the Difference Principle, the sense in which his view was fair to people with different conceptions of what is good, and he clarified the "Kantian interpretation" of justice as fairness (see 1999: chs 11–15). In 1980 Rawls published "Kantian Constructivism in Moral Theory" (the Dewey Lectures), in which he carefully described the details of the contract so that it represented the Kantian idea of free and equal agents who are rational and reasonable. On this constructivist view, there is no claim that the moral principles are "true" or represent a prior moral order.

Beginning in the mid-1980s, Rawls became concerned that he had underestimated the importance of the divergence among comprehensive moral and religious views that would emerge under the very conditions of liberty promoted by his theory. Could people with such divergent views have a stable agreement on a conception of justice? To answer this question and to accommodate the "reasonable pluralism" he thought unavoidable, Rawls revised his account of stability and political justification in papers leading up to and including *Political Liberalism* (1993). In *Liberalism*, he replaced Kantian constructivism with "political constructivism," and the same ideas about free and equal citizens are used to construct a political conception of justice, again with no explicit claims about moral truth.

Rawls's last major work is thus motivated by a central question about justification that evolved from his thesis and first publication: How can reasonable people with divergent moral and religious views come to agree on and abide by fair terms of cooperation?

Justice as fairness

The social contract

Rawls revives the social contract as a way to specify fair terms of social cooperation in the form of a hypothetical, not an actual or historical, agreement. The appeal to a contract embodies three main ideas. First, it is a form of *procedural* justice. When we do not have a prior principled agreement on what counts as fair or just, for example, about

how to reconcile concerns about liberty and equality and efficiency, we must rely on a procedure that is fair to all parties. We then can count the outcome of that procedure as fair or just.

Second, a fair procedure must embody features that are reasonable in light of the nature of the problem it addresses. In this case, the problem is to find principles of justice that "free and equal" persons can all agree provide the basis for a "well-ordered society." As citizens or persons, we are free in the sense that we can form and revise a rational plan of life that specifies our conception of the good. We are equal in that we all normally have an adequate sense of justice, a disposition to seek and abide by terms of fair cooperation. A well-ordered society is one in which citizens accept and know that others accept the same principles of justice, and those principles govern its basic institutions.

Third, if properly designed, the contract situation represents an Archimedean Point. It stands outside the biased or self-interested beliefs we happen to hold, as well as the entrenched inequalities that may motivate them. From this standpoint, the contractors can leverage new agreements on points of controversy by building on relatively fixed or uncontroversial points.

Though we might understand that an actual contract binds those who make it, or those who implicitly consent to it, why think Rawls's hypothetical contract tells us anything about what we ought to do? The answer must be that we share enough substantive moral agreement about our nature as free and equal persons, the goal of arriving at a well-ordered society, and the appropriateness of the design of the contract situation that we accept it as a procedural solution to the problem of justice.

The original position

The hypothetical contract involves "reasonable" constraints on contractors who must make a "rational choice" of principles of justice, presented as pair-wise comparisons. The contractors are limited in both knowledge and motivations, and thus they must not be confused with the "fully informed" bargainers or rational choosers who populate standard rational choice problems. They operate behind a "thick" veil of ignorance that blinds them to information about their age, race, gender, class position, the society they will enter and its position in history. They are also blind to their "rational plan of life" of conception of the good, including their system of moral and religious values. This thick veil assures that their choice of principles will not be affected by the self-interest that might come from knowledge of any of these facts about themselves. Contractors do, however, have general social knowledge or they would not be able to evaluate the choice of principles for their effects on well-being. Their motivations are also constrained. They are "mutually disinterested," meaning they are concerned about their own well-being and the well-being of those in a generation or two either side, but they are not generally benevolent, malevolent, or envious.

Having blinded contractors to their own detailed or "thick" conception of the good, Rawls must provide them with a basis for determining how one set of principles will make them better off than another. Otherwise there is no basis for a rational choice of principles. Rawls introduces a set of "primary social goods," containing rights and liberties, powers and opportunity, income and wealth, and the social bases of self-respect.

A weighted index of these objectively measurable goods is the basis for measuring the effects of alternative principles on well-being and for measuring inequalities between (representative members of) social groups. Consequently, the rational choice problem facing contractors is to decide which of two alternative principles under consideration leads to the highest index score for them.

Together with the requirement that all contractors have veto power over choices, the result of the "reasonable" constraints is to establish a baseline of equality that eliminates the influence of entrenched social inequalities.

Principles of justice

Rawls argues that his contractors would choose three principles of justice in preference to utilitarianism. His First Principle assures citizens they will have a set of equal basic liberties, including freedom of thought, expression, and association, security of the person, and rights of political participation. Rational contractors know that they may have fundamental moral and religious commitments, even if behind the veil they do not know exactly what they are. Once they reach some modest threshold of material well-being, having the liberty to pursue those commitments is not something they would trade for increments in income and wealth. Recognizing others and being recognized by them as political equals is an important social basis of self-respect as well.

So important are the recognitional aspects of the effective exercise of political participation rights that Rawls argues for special institutional protections of them. These institutional protections, such as public funding of elections, are intended to make sure that political participation rights are not merely formal but actually effectively exercisable by all, regardless of other inequalities. For these reasons, contractors would assure themselves these basic liberties directly through the First Principle rather than depending for them on the outcome of an uncertain utilitarian calculation.

Rawls's Second Principle actually consists of two other principles. The fair equality of opportunity principle not only prohibits legal and quasi-legal barriers to opportunity, as would the weaker "formal" equality of opportunity principle, but also requires that positive steps be taken to mitigate the effects of social and economic contingencies on the developments of the talents and skills. These steps minimally include measures such as the provision of public education, but further early childhood interventions and family supports, such as day care, might be necessary to support fair equality of opportunity for both children and women. Both the First Principle and the fair equality of opportunity principle require certain kinds of equality. They are given priority over the Difference Principle, which allows certain inequalities.

The Difference Principle requires that inequalities, for example in income or wealth, be allowed only if they work to make the worst-off groups (and then the next worst off, etc.) as well off as possible. The idea is that it would be irrational for contractors to insist on equal shares of a small social produce if incentives would create a larger social product that could be divided so that all, even the worst off, benefit by getting more (according to the index of primary goods) than they would without incentives and inequalities. So far, this argument establishes only that it is irrational to disallow inequalities that advantage all. The Difference Principle, however, is very demanding

in that it requires inequalities to make the worst off as well off as possible (it is much more than "trickle down"). Crucial to Rawls's argument for the Difference Principle is his claim that the very high stakes (lifetime prospects of well-being) and the great uncertainty imposed by the veil of ignorance mean that a "maximin" (maximize the minimum) principle is required as a principle of rational choice. Contractors are not permitted to gamble that they have an equal chance of being in any social position, a gamble that would make the principle of average utility preferable to them. Rather, the maximin principle requires they maximally protect the worst off through the Difference Principle. To make the Difference Principle seem less odd, Rawls also argues intuitively that both it and the fair equality of opportunity principle work to mitigate the effects of morally arbitrary social contingencies.

Utilitarians, Rawls notes, believe that in pursuing the aggregate welfare, the advantages of some outweigh the losses of others, much as the expenditure of effort in acquiring skills at one stage of life will be offset by greater rewards at another. Rawls's principles together better recognize the "separateness of persons." They afford stronger protections to individuals so that the advantages of social cooperation work more directly to the benefit of all.

Basic structure

Rawls intends his principles of justice to regulate the basic structure of society, that is, those major social institutions, such as the political constitution and the principal economic and social arrangements, that have "profound" effects on people because they distribute basic rights and duties and determine the division of advantages from social cooperation. In his later work, he explicitly includes the family in the basic structure. The principles of justice do not directly apply to the relationships individuals have with each other or in private associations. Rawls suggests there is an important division of moral responsibility: society assures that citizens' needs are met through the principles of justice, which regulate basic institutions. Individuals are responsible for pursuing their rational plans of life within the constraints imposed by justice.

Primary social goods

Rawls's rejection of utilitarian measures of well-being, such as welfare or desire satisfaction, in favor of an index of primary social goods, was challenged on several grounds. The index seems incomplete, for it fails to tell us who is worse off, the rich but ill person or the poor but well one. More generally, individual variations, such as those caused by disease or disability, would mean that individuals with the same primary social goods would actually have quite different capabilities. Some conclude that Rawls's focus on the "resources" included in the index of primary social goods means he is concerned about the wrong "space"; egalitarians should focus more directly on individuals' "opportunity for welfare or advantage" or on the capabilities or positive freedom they have.

In *Theory*, Rawls had made the simplifying assumption that all individuals were fully functional over a normal lifespan. This assumption invited these objections: that the index of primary goods was insensitive to important individual differences, such as disease or disability. By viewing disease and disability as impairments of the range

of opportunities open to people, it is possible to extend Rawls's theory to include problems of health and disease, and Rawls endorses such an extension in his later writings. Quite surprisingly, Rawls's three principles then constitute a fair distribution of the major "social determinants" of health, according to current work in the social sciences. In his later writings, Rawls also replies to the objection that the primary social goods might not be valued in the same way by people holding quite different conceptions of the good. He reformulates them as crucial "all purpose means" for meeting the needs of citizens. This reformulation, together with the extension to health care, makes his view converge more with that of some of his critics, for it suggests that justice as fairness is aimed at guaranteeing that all citizens' "needs" – as citizens – are met and that therefore they all have the capabilities to function – as citizens – as free and equal.

Wide reflective equilibrium

The principles that contractors choose must match our considered judgments about what is just in "reflective equilibrium." To achieve "reflective equilibrium," we work back and forth between our considered judgments about particular instances or cases, the principles that govern them, and the theoretical considerations that bear on accepting these considered judgments or principles, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them. For Rawls, this means we should revise the constraints on choice in the Original Position until we arrive at a contract that yields principles that are in reflective equilibrium with our considered judgments. Thus, in his early work, the method of reflective equilibrium plays a role in both the construction and justification of his theory of justice.

There seems to be little justificatory force to achieving coherence or reflective equilibrium solely among principles and judgments about particular cases. Unless we think we have special knowledge of either the principles or judgments, which Rawls does not, such a *narrow* reflective equilibrium captures only what we happen to think is just. It does not show us that we are justified in holding those particular beliefs. Rawls believes we have no better method of justification than seeking a *wide* reflective equilibrium. This method broadens the field of relevant moral and non-moral beliefs to include both an account of the conditions under which it would be fair for reasonable people to choose among competing principles, and evidence that the resulting principles constitute a feasible or stable conception of justice, that is, that people could sustain their commitment to such principles. Our beliefs about justice are justified (and, by extension, we are justified in holding them) if they cohere in such a wide reflective equilibrium.

Central to the method of reflective equilibrium is the claim that our considered moral judgments about particular cases carry weight, if only initial weight, in seeking justification. Vigorous criticism of this claim comes from utilitarians, who denounce such "intuitions" as the results of historical accident and bias. Since, however, utilitarians allow individuals' desires or preferences to count in calculating what is good and right, and these desires are also affected by historical accident and bias, Rawls's openness about exposing moral judgments to comprehensive criticism in reflective equilibrium may be less harmful than the utilitarian approach.

In *A Theory of Justice*, Rawls seemed to think that all people might converge on a common or *shared* wide reflective equilibrium that included "justice as fairness." We would be led to that equilibrium by philosophical argument about the various moral beliefs that contribute to the social contract approach, the details of the Original Position, and the arguments made within it. In his later work, Rawls modifies this view (see "Justification revisited" below).

Stability and feasibility

Principles of justice must not only be chosen by contractors and match our judgments in reflective equilibrium, but they must be more stable than alternative views. People raised in a well-ordered society must find that conforming to them involves less strain of commitment than conformity with alternatives. For example, the worst off arguably would find the strain of commitment less under the Difference Principle, which makes them as well off as possible, than they would under a utilitarian principle that simply maximized aggregate or average utility, assuming benefits to others compensated them for their losses. Because the autonomy exercised enabled by the principles of justice would be viewed as a good by people, Rawls thought his view stable. A growing respect for pluralism led Rawls to revise this argument for stability.

Justice as political

Burdens of judgment and reasonable pluralism

Reasonable people, especially under conditions in which they enjoy basic liberties, will tend to develop divergent comprehensive philosophical and religious views through which they assess what is valuable in life. By "reasonable," Rawls means people who are concerned to live with others on fair terms, assuming that the others are so willing. Reasonable people also understand that to be fair the terms of cooperation must be ones that other free and equal persons can accept. Reasonable people will recognize that disagreements arise among them because of the "burdens of judgment." These burdens include the conflicting and complex evidence that bears on issues, the disagreements about how to weight considerations, the vagueness of some of our concepts, the effects of the totality of a person's experience on how she weights considerations, the multiplicity of normative considerations that are relevant and from which a selection must be made in any specific case. We are driven, Rawls concludes, to accept reasonable pluralism about many matters of importance. This is a basic fact of political life, and even among reasonable people we will find disagreements that threaten the original suggestion that philosophical argument could produce convergence on the same wide reflective equilibrium.

Overlapping consensus

Rawls addresses the problem of producing stable agreement despite reasonable pluralism by recasting justice as fairness as a "free-standing" *political* conception of justice. The key ideas out of which justice as fairness (or other, alternative reasonable

political conceptions of justice) are constructed, for example, the idea that citizens are free and equal, are now taken to be shared elements of our political life, that is, of a public, democratic culture. These ideas are already held or accepted by most people who share that culture, whatever other views they diverge on. In effect, it is not philosophy alone – aided by universal reason – that has led people to converge on these ideas, but a shared set of institutions and history. The appeal to a shared democratic culture, however, is not a concession to the "communitarian" critics of Rawls, who had complained that a shared conception of the good must unite people and form the basis for justice; instead, it is a way for Rawls to seek agreement among those who disagree about such views of the good, among other things.

Rawls suggests that we think of the political conception of justice as fairness as a "module" with its own internal principles, reasons, and standards of evidence. For example, justice as fairness includes the two principles of justice ordered in a particular way. Together these ordered principles, illuminated by the shared background ideas and publicly defendable standards of evidence and reasoning, specify the content of "public reason" as it is used to deliberate about matters of justice. This module should be complete: it should give "reasonable" answers to a broad range of questions about "constitutional essentials and basic questions of justice."

These answers are "reasonable," however, in light of the kinds of reasons to which the political conception is restricted. In effect, the justification for these answers only goes so far. It appeals only to reasons contained in the public view. Rawls calls this "*pro tanto* justification."

People with divergent comprehensive moral and religious views can overlap in their acceptance of a conception of justice, the most reasonable of which Rawls thinks is justice as fairness. He draws an analogy to the same theorem's being provable within different axiomatic systems. Nothing that turns on the comprehensive views plays a role in the public justification of the module. No claims about moral truth and no specific moral or philosophical views that are the distinctive features of such comprehensive views play a role.

Overlapping consensus is not a compromise or *modus vivendi* among competing groups that hold different moral conceptions. Public justification of the view must be for the "right reasons" and turn on the acceptability of the module and the ideas it rests on to those who hold those views.

Public reason

An idea that becomes central in Rawls's later work is that of "public reason." In *Political Liberalism*, Rawls argues for a rather restrictive view of public reason, attempting to restrict the introduction of religious and other views into public debate, especially by public officials and even in the thinking of citizens as they vote. This view was widely criticized, and Rawls adopts a more relaxed, "wide" view of public reason in his last paper on the topic, "The Idea of Public Reason Revisited." In the wide view, deliberations about justice, especially by public officials, are governed by a "proviso." Reasonable comprehensive moral or religious doctrines may be introduced into public political discussion at any time, and there may be good reasons for doing so, *provided*

that proper political reasons are also offered that are sufficient to establish the same conclusions. This proviso applies to public political culture and not to debate in the background culture, which has no such restriction on it.

Justification revisited

To say that a claim about what is just is justified solely by public reason (or *pro tanto*) is not yet to say that it is a fully justified belief for a particular person. The criterion for full justification ultimately remains acceptability in wide reflective equilibrium, and *pro tanto* justification deliberately refrains from seeking such deeper justification. By not seeking or alluding to deeper justification, *pro tanto* justification does not alienate those who have different reasons for accepting the module.

We obtain the greatest stability we can for a political conception of justice, Rawls argues, answering his central question in *Liberalism*, when there is the right type of "overlapping consensus" on it, that is, when there is overlapping consensus for the right reasons. People with different comprehensive moral views must justify for themselves, by their own lights, that is, in their own wide reflective equilibria, the acceptability of the module. Their rationales will thus differ in ways that reflect their other philosophical, moral, and religious beliefs. Some may insist, for example, that there is "moral truth," others deny it. Some might see the principles of justice as forms of divinely given natural law; others may see it as a human construction. Ultimately, people are justified in accepting justice as fairness if it is acceptable to them in the different wide reflective equilibria they can achieve.

If there is *general* acceptance in this way of the module within the different "reasonable" comprehensive views in a society, Rawls says that we have "general" reflective equilibrium. General reflective equilibrium is not a shared wide reflective equilibrium – except for the overlap on the module.

Current applications and controversies

At the beginning of the twenty-first century, Rawls's work continues to stimulate extensive discussion in several very active fields of political philosophy, including the following:

The family and feminism

In his latest writings, Rawls emphasizes that the family should be included in the basic structure and thus be regulated by principles of justice. At the same time, concerns about equality must be reconciled with concerns about the liberty of families to pursue religious or moral views that involve gender role differentiation affecting mothers and children. Rawls imagines robust institutional protections of women, including day care and other family support systems; at the same time, he imagines the debate about gender roles to be carried on in the background culture and not through intrusions into the family. Nevertheless, Rawls's emphasis on principles of justice should be contrasted with virtue-based feminist approaches to ethics.

Egalitarianism

A substantial body of egalitarian literature has arisen that challenges Rawls from various directions. One prominent view suggests that Rawls's intuitive arguments for "democratic equality," which appeal to the moral arbitrariness of social and natural contingencies, yield a more egalitarian view than is embodied in Rawls's principles. On this view, we are owed compensation for any deficit in opportunity for welfare or advantage that arises through no fault or choice of our own. Another egalitarian challenge is that the leeway Rawls allows to individuals to pursue incentives or to make selfish domestic choices will undermine the possibility of achieving optimal results, as judged by Rawls's own principles.

International justice

An early challenge to Rawls was that he failed to discuss what kinds of obligations of justice are owed across national boundaries; a related criticism is that his theory is too much wed to the idea of nation states. Rawls's late publications on "The Law of Peoples" extend his contractarian views to discuss human rights.

Democratic deliberation

The view of democracy that emerges in Rawls emphasizes political participation as a way of realizing our moral capabilities. Democracy is not simply a procedural method of achieving agreement that we must employ because we lack substantive agreement on various matters. Instead, Rawls provides a foundation for emphasizing the deliberative elements of democratic theory.

Bibliography

Works by Rawls

1971: *A Theory of Justice*, Cambridge, MA: Harvard University Press.

1993: *Political Liberalism*, New York: Columbia University Press. (Paperback edition (1996) contains a new introduction and "Reply to Habermas," from *Journal of Philosophy* 92/3 (March 1995).)

1999: *Collected Papers*, ed. S. Freeman, Cambridge, MA: Harvard University Press.

Works by other authors

Daniels, N. (ed.) (1975) *Reading Rawls*, New York: Basic Books.

Freeman, S. (ed.) (forthcoming) *Companion to Rawls*.

Reath, A., Herman, B., and Korsgaard, C. M. (eds.) (1997) *Reclaiming the History of Ethics: Essays for John Rawls*, New York: Cambridge University Press.

Symposium on Rawlsian Theory of Justice: Recent Developments, *Ethics* 99/4 (July 1989).

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

30

Thomas S. Kuhn (1922–1996)

RICHARD GRANDY

Thomas S. Kuhn's second monograph, *The Structure of Scientific Revolutions* (1962) is the most widely read and most influential book on the philosophy of science of the twentieth century. It spawned the ubiquitous use of the term "paradigm" in popular culture, including cartoons and business management courses, and a million copies have been sold in almost twenty languages.

The central thesis of the book is that the nature of scientific development had been seriously misunderstood by philosophers and scientists, and that, in the words of the opening sentence: "History, if viewed as a repository for more than anecdote or chronology, could produce a decisive transformation in the image of science by which we are now possessed." The image he sought to transform was one in which science is cumulative, varying in the speed of its progress, but always moving forward, an image in which scientific controversies are a small and unimportant part of the process, friction in the wheels of progress.

The contrasting image he championed portrays mature sciences as alternating between two kinds of change. The first are periods of cumulative progress in which scientists apply generally accepted theories to the unresolved questions in a domain according to a shared understanding of what constitutes a reasonable scientific question and of what criteria are used to judge answers. This "normal science" is a very sophisticated form of puzzle-solving and can require great ingenuity, but occurs within a stable framework of tradition. In contrast, the alternating periods of "revolutionary science" consist of confrontation between two diverse understandings of what constitutes a reasonable question and what criteria should be used to adjudicate disputes. In *The Structure of Scientific Revolutions* Kuhn used the term "paradigm" to both define and explain the difference between the two kinds of science: normal science consists of the elaboration of an accepted paradigm, while revolutionary science consists of the overthrow, or attempt to overthrow, an accepted paradigm. In addition, the notion of paradigm played an essential role in distinguishing prescientific preludes to a science, for example optics before Newton, because the critical step in making the transition to a science consisted of convergence by a scientific community on a paradigm.

Reactions to the book by philosophers and natural scientists were numerous, vociferous, and almost all negative. Some critics said that Kuhn made most of science – normal science – seem pedestrian and almost unnecessary, in spite of his clear

insistence that it was only the persistent pursuit of puzzles by first-rate minds that would eventually generate the anomalies which would lead to revolutionary science. But the majority of critics focused instead on the account of revolutionary science; according to many, the processes of revolutionary change as described by Kuhn constituted irrational mob rule and were antithetical to the view of science as the epitome of reason. Much of the attention centered on the claim that opposing paradigms are incommensurable, that the meanings and often the referents of the terms of the theories differ so that no direct simple comparison between them is possible. The reaction among social scientists was more mixed; some embraced the central themes and became obsessed with whether their field had yet completed its preparadigmatic preparation for sciencehood.

The analytic apparatus of the book, especially the central notion of a paradigm, came in for particularly severe scrutiny. One reviewer discerned twenty-two distinguishable senses of "paradigm" in the text. Kuhn was astounded at what he saw as widespread misunderstanding and misrepresentation of his ideas, but recognized the need to clarify the central notion of paradigm and related apparatus. The first fruition of this rethinking appeared in a number of papers during the late 1960s and in a "Postscript," which was published in the second edition of *Structure* in 1970. Before elaborating on the modification in the "Postscript," it will be useful to sketch some of the path by which he reached the views behind *Structure*.

Kuhn's status as a philosopher is difficult to assess because his training, career, and indeed the nature of his influence are very unusual. Thomas S. Kuhn was born in Cincinnati, Ohio in 1922 and received his Bachelor's (1943), Master's (1946) and doctoral degrees (1949) from Harvard University in physics. He only began to read seriously in the history of science when asked by James B. Conant, then President of Harvard, to assist in preparing a historically oriented undergraduate science course for non-science students.

A pivotal moment occurred in 1947 while he was reading Aristotle, trying to ascertain how much mechanics Aristotle understood. His conclusion was that Aristotle understood little or no mechanics and indeed seemed to be a poor observer and unsystematic scientist. He was puzzled by how one of the greatest intellects in the history of western thought could have been so confused. Then, suddenly, "the fragments in my head sorted themselves out in a new way, and fell into place together. My jaw dropped, for all at once Aristotle seemed a very good physicist indeed, but of a sort I'd never dreamed possible" (Thalheimer lecture p. 32). Among the pieces that had sorted themselves out was the insight that for Aristotle, the Greek expression that is translated as "motion" means not only a change of location, but any of a wide variety of changes, of which change of location is only one. Looking at the world in this new way with this transformed vocabulary, Aristotelian mechanics made very good sense of many observations, albeit many of those observations would not be regarded as relevant to modern mechanics.

After completing his dissertation in physics, he spent three years as a member of the Harvard Society of Fellows broadening his historical and philosophical knowledge. Then ensued an appointment teaching history of science at Berkeley in which much of his time, by his own observation, was spent preparing lectures in a field in which he had no formal training. In 1957 he published *The Copernican Revolution*, a well-received

account of the conceptual and technical obstacles to making the transition from a geocentric to a heliocentric universe. The central ideas of *Structure* are discernible in this first book, but the claims are much narrower and generally less philosophical. There is considerable focus on the idea that the transitions from the Aristotelian–Ptolemaic universe to the Copernican–Galilean–Newtonian one are not transitions that can be arrived at by small incremental steps. To see the universe as centered on the sun, not on the earth, can only be accomplished as a dramatic change. But no wider claims are made in the first book about how widespread this kind of transformation has been in the history of science.

Structure represented the generalization of that idea to the larger canvas of the physical sciences generally. The preface to *Structure* indicates in some detail the extent to which he is aware that there are serious gaps and shortcomings in the philosophical development of key concepts. However, he had contracted to produce a monograph within fairly severe size limits and the editors were pressing him to complete the manuscript.

One little known ironical aspect of the publication of the book is that although the logical positivist conception of science is a primary target of Kuhn's criticisms, the monograph was first published as Volume II, no. 2, of *The Encyclopedia of Unified Science*, the publishing organ of the logical positivist movement. The editors responsible for soliciting and encouraging the manuscript, Charles Morris and Rudolph Carnap, were enthusiastic about the monograph and its importance for philosophy of science. The second edition of the book indicated its status as part of the *Encyclopedia* less saliently, and by the third edition in 1996 no mention is made of the original imprimatur.

In 1964 Kuhn moved from Berkeley to Princeton, becoming a member of the history department but also joining the graduate program in the history and philosophy of science. In seminars there, as well as in lectures and correspondence, he revised and clarified the ideas of *Structure*. In particular, the use of "paradigm" was to be replaced by either "disciplinary matrix" or "exemplar," thus recognizing a major two-fold ambiguity in the original term. A disciplinary matrix consists of symbolic generalizations, metaphysical assumptions, models, values, instruments, and exemplars. Thus a disciplinary matrix is a constellation of elements which define a world-view and characterize a scientific community. Since a disciplinary matrix contains many elements, there can be varying degrees of congruence among members of a community.

The symbolic generalizations are the most familiar element; these would be equations such as Newton's Laws or Boyle's Law. Metaphysical assumptions concern the basic elements of the universe; examples would be the assumption that a vacuum is impossible, that action-at-a-distance is impossible, that the universe is governed by deterministic laws, and that all matter consists of atoms in a void. Models are easier to illustrate than describe: the model of the atom as a miniature solar system, the model of a gas as a collection of a large number of very small particles in rapid motion, heat as a fluid, and so on. Values include simplicity, generality or scope, accuracy, reproducibility of results.

Exemplars, which are both an element of disciplinary matrices, but also a significant second sense of paradigm, are examples of notable scientific accomplishment which set a standard for future researchers. For example, the rigor of Euclid's geometry was an exemplar for many disciplines, and the number of fields that have proclaimed

Copernican revolutions is legion. More recent examples might be the predicted discovery of Uranus or the discovery of the double helical structure of DNA. The exemplars provide a glue for the elements of a disciplinary matrix by bringing together the examples in concrete accomplishments. It is an essential part of the Kuhnian picture that the examples can be extended in various ways, so that the exemplars provide guidance but not rules.

A major emphasis in Kuhn's discussion of scientific change was the sudden and involuntary transformation of perception and belief. This clearly originated in his own experience in understanding Aristotle and he illustrated it by giving examples of Gestalt switch figures in his book, for example, a line drawing which can be seen either as an old woman from one perspective or a young woman from another. These ideas biographically stemmed back to Kuhn's experience in 1947, but others had also been struck by similar ideas, and an articulate presentation of them could be found in N. R. Hanson's *Patterns of Scientific Discovery* four years before the appearance of *Structure*, which refers approvingly to Hanson in a number of places.

Kuhn was particularly perplexed and frustrated by the accusation that he was undermining the rationality of science. He strongly believed that science is an epitome of rationality, and thus the processes involved in the development of science, including both normal and revolutionary science, must be essential ingredients in the rationality of science. His goal in overthrowing the accepted image of scientific processes was to cast aside a false understanding of rationality and to begin the process of replacing it with a more sophisticated and historically accurate apprehension.

He made one strategic decision in completing the manuscript of *Structure* and publishing his ideas in that abbreviated and highly unfinished form. He made a second decision in the late 1960s to publish the "Postscript" at the end of the second edition of *Structure*, rather than attempting a thorough revision that would systematically replace the occurrences of the ambiguous "paradigm" with the appropriate term from the vocabulary of disciplinary matrices and exemplars. This meant that even after 1970 new readers of *Structure* became aware of the extensive terminological and conceptual changes only after reading the original 170-page text and being thoroughly immersed in the sweeping and ambiguous vocabulary of "paradigms."

This decision was the result of Kuhn's recognition that reworking *Structure* was not a very good option since he was still in the midst of changing his views, and so the "Postscript" strategy was a stopgap until he could reach the stage where a new and more thorough book was prepared. During the 1960s and 1970s he gave frequent graduate seminars on *Structure* and his further thoughts, as well as giving lectures and publishing intermediate hints of his elaborations. Two major influences on his thinking were conferences in London in 1965 and Champaign, Illinois in 1969, at which *Structure* was a major critical focus. The proceedings of these were eventually published as *Criticism and the Growth of Knowledge* (edited by Lakatos and Musgrave) and *The Structure of Scientific Theories* (edited by Suppe). In 1977 he published *The Essential Tension*, a collection of his essays ranging from reprintings of pre-*Structure* papers to items that appeared for the first time in that volume. The essential tension referred to is that between the desire to assimilate all data and observations within the current paradigm and the desire to find revolutionary new solutions.

He continued to be heavily involved in history of science, the main culmination of which was the publication in 1978 of *Black-Body Theory and the Quantum Discontinuity, 1894–1912*. In 1979 he left Princeton for the MIT department of philosophy and linguistics, where he became a professor of philosophy for the first time. Subsequently his research focused more exclusively on refining his answers to the questions raised about *Structure*: about the nature of incommensurability, the relation between disciplinary matrices and scientific communities, the elements of disciplinary matrices, rationality, and theory choice. His analytic tools also shifted; the "Postscript" was phrased in terms familiar to readers of Quine's *Word and Object*, whereas his later work invoked possible worlds (see LEWIS) and rigid designators (see KRIPKE).

One recurring issue was the clarification of his ontological views. Probably the most infamous sentence of *Structure* occurs on p. 150: "In a sense I am unable to explicate further, the proponents of competing paradigms practice their trades in different worlds." Well disposed critics urged that he probably did not really mean to say that they were in different worlds, just that the world looked very different to them. But he was adamant that there was an important insight in the stronger claim. This was important to him because, for instance, he also wanted to claim that before the medieval paradigm change that introduced the concept of the pendulum, there were no pendulums but only swinging stones (p. 120). His attempts to clarify this and related locutions led him to further investigations of the interrelations of language, concepts, and perception and to propose that these were at least partially constitutive of the world.

He became an emeritus professor in 1989 but rather than diminishing his efforts, he used this as an opportunity to spend more time on his research agenda. At the time of his death in 1996 the solutions were still not in his grasp and the envisioned conclusive manuscript was still in an early stage. An extensive study of his later work is *Reconstructing Scientific Revolutions*, a 1993 translation of Paul Hoyningen-Huene's 1989 book. Hoyningen-Huene worked closely with Kuhn in producing the book and it is almost a collaboration. A thorough evaluation of Kuhn's work can be found in *World Changes*, edited by Paul Horwich, which is the revised product of a 1990 conference on Kuhn's work and includes responses to his critics.

I have waxed biographical to underline the peculiarity of evaluating Kuhn from the context of analytic philosophy. He had no formal training in philosophy, and his most influential work was completed before he was very thoroughly conversant with the intricacies of the analytic tradition. But he was already sufficiently familiar with it at the time of the writing of *Structure* to recognize that he would be accused of confusing the context of scientific discovery with the context of scientific justification, a distinction formalized by Reichenbach but which was widespread in the tradition before his articulation. Kuhn's response to the accusation was to question the distinction: to argue that only a historically inaccurate and oversimplified view of scientific development would permit such a distinction and that to maintain such a distinction was to doom epistemology to sterility.

Other philosophers of science – Hanson, Toulmin, Feyerabend, Hesse, among others – published books with at least similar themes in the late 1950s and 1960s, but none of those had the same effect or, possibly excepting Feyerabend's *Against Method*, produced so strong a reaction in readers and reviewers.

It would be easy to underestimate the influence *Structure* and Kuhn's subsequent work had directly on philosophy of science and indirectly on analytic philosophy generally. A large percentage of a generation of philosophers of science spent a considerable portion of their careers showing that Kuhn was wrong – wrong about incommensurability, wrong about paradigms, wrong about the role of scientific communities, wrong about rationality, wrong about the relevance of psychology for philosophy of science, and most significantly, wrong about the import of history of science for philosophy of science.

However, the results of the inquiries demonstrating the defects and errors of *Structure* bear a far greater resemblance to *Structure* than to its predecessors. The situation seems comparable to the role of Piaget in developmental psychology. Few, if any, of Piaget's specific claims about developmental stages or even about the abilities (and inabilities) of children at various ages have withstood further more sophisticated research. But Piaget brought the field into existence and without his impetus it is not clear that any of the further research would have been done.

One could argue that I have overstated the impact of Kuhn's work; other philosophers, including Carnap and Hempel, as well as the previously mentioned authors, were calling, albeit more quietly, for a rethinking of the image of science that had been dominating philosophy of science (see CARNAP and HEMPEL). The received view of scientific theories was under attack both from those who questioned the pivotal distinction between theoretical and observational vocabulary, but also from the structuralist approach to theories championed by Braithwaite, Suppes, van Fraassen, and Suppe. On the other hand, the most refined version of structuralist theories, that produced by Stegmuller, Sneed, and others, drew strong inspiration from Kuhn.

The importance of careful historical case studies, of consideration of the broader context of scientific developments, of the cognitive abilities and constraints on scientists, of the "external" influences such as motivation and competition, are all now taken for granted as part of philosophy of science. Debate rages about the relative importance, interpretations, and so on, but in the background there are shared assumptions that were not in place before the influence of Kuhn's work. Examples of important recent books that are not always cognizant of their Kuhnian heritage, but which can be seen to be following in a Kuhnian tradition are Longino's *Science as Social Knowledge*, Giere's *Explaining Science*, and Kitcher's *The Advancement of Science.* His work has also inspired the development of historicist, feminist, and sociological movements in the philosophy of science with whose doctrines he often disagreed.

Bibliography

Works by Kuhn

1957: *The Copernican Revolution*, Cambridge, MA: Harvard University Press; 7th edn., 1985.
1962: *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press. (2nd edn., with "Postscript," 1970.)
1977: *The Essential Tension*, Chicago: University of Chicago Press.
1978: *Black-Body Theory and the Quantum Discontinuity, 1894–1912*, New York: Oxford University Press.

Works by other authors

Hoyningen-Huene, P. (1993) *Reconstructing Scientific Revolutions,* trans. Alexander Levine, Chicago: University of Chicago Press.

Horwich, P. (ed.) (1993) *World Changes: Thomas Kuhn and the Nature of Science*, Cambridge, MA: MIT Press.

Lakatos, I. and Musgrave, A. (eds.) (1974) *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press; 3rd edn., corrected, 1977.

Suppe, F. (ed.) (1974) *The Structure of Scientific Theories*, Urbana, IL: University of Illinois; 2nd edn., 1977.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

31

Michael Dummett (1925–)

ALEXANDER MILLER

Michael Dummett (Wykeham Professor of Logic at Oxford 1979–92) is one of the most important and influential British philosophers of the second half of the twentieth century. In addition to making seminal contributions to the exposition and study of the philosophy of Frege, Dummett started a debate – concerning how issues in metaphysics might best be prosecuted via arguments in the philosophy of language and theory of meaning – which continues to be one of the central issues in contemporary analytic philosophy. The two are intimately related. We are given a large-scale exposition and partial defense of a broadly Fregean theory of meaning. It is then argued that the realist position in metaphysical debates about a disputed subject matter is best cast as a semantical thesis about the meaning of sentences concerning that subject matter. Once Wittgensteinian insights about linguistic understanding and language mastery are incorporated into the Fregean theory of meaning, it emerges that the semantical thesis in which the realist view is best cast turns out to face very serious challenges. An anti-realist alternative is explored, drawing on the theory of meaning proposed by intuitionism for mathematical statements, and it is argued that one consequence of this is the rejection of certain theorems of classical logic, such as the law of excluded middle.

Frege

Dummett's exposition and partial defence of a Fregean theory of meaning is writ large throughout his work, but the key texts are *Frege: Philosophy of Language*, *Truth and Other Enigmas* (essays 7–9), *The Interpretation of Frege's Philosophy*, *Frege and Other Philosophers*, and *Origins of Analytic Philosophy*.

For Frege, whether or not a sentence is grammatically well formed is determined by syntactical rules. These are the province of *syntax*, and tell us how expressions from different syntactic categories may be combined to form grammatical sentences. In the province of *semantics*, the *Bedeutung* of any expression is that feature of it that determines whether sentences in which it occurs are true or false. The *Bedeutung* of a sentence is its truth-value (true or false). Whether a sentence is true is determined by the *Bedeutungen* of its constituents. Expressions from different syntactic categories are assigned different types of *Bedeutung*: the *Bedeutung* of a proper name is the object

it stands for, the *Bedeutung* of a predicate is a first-level function from objects to truth-values, the *Bedeutung* of a sentential connective is a first-level function from truth-values to truth-values, and the *Bedeutung* of a quantifier is a second-level function from concepts (first-level functions) to truth-values. The *Bedeutung* of a constituent of a sentence is determined by its *Sinn* or sense (that ingredient of its meaning that determines its contribution to the truth or falsity of sentences in which it may appear), which in turn determines, in conjunction with the senses of the other constituents, the *Sinn* of the sentence. The *Sinn* of a sentence is the thought which it expresses, conceived not as some psychological episode or entity, but as a *truth-condition*: the condition which must obtain if the sentence is to be true. The *Sinn* of an expression is what someone who understands an expression grasps: our understanding of whole sentences therefore consists in part in our grasp of their truth-conditions (see FREGE).

Much of Dummett's work consists in a sophisticated elaboration of the theory thus crudely summarized, and an examination of the other notions – such as *force* (that ingredient in meaning which distinguishes, e.g., assertions from questions and commands) and *tone* (that ingredient in the meaning of, e.g., "and" which distinguishes it from "but" even though it has the same *Sinn*) – that need to be added to the notions of *Sinn* and *Bedeutung* in order to obtain a comprehensive theory of meaning. *Inter alia*, Dummett defends the Fregean theory in the face of attacks from the causal theory of reference advocated by Kripke (1973: appendix to ch. 5) and the holistic picture of language advanced by Quine (1978: essays 9 and 22, 1973: ch. 17). On Dummett's interpretation of Frege, it is possible for an expression, such as the proper name "Vulcan," to have a *Sinn*, even though it has no *Bedeutung*, since there is no object for which it stands: sentences containing expressions which have no *Bedeutung* themselves fail to have a *Bedeutung*, that is, fail to possess a truth-value (1973: ch. 6). This facet of Dummett's interpretation is challenged in important work by Gareth Evans (1982) and John McDowell (1998a: essays 8–12). Whereas for Dummett, the sense of an expression is a method or procedure for determining its *Bedeutung*, so that the sense of a sentence, for example, is a method or procedure for determining its truth-value, a method which can exist even if the sentence in question has no truth-value, for Evans and McDowell the sense of an expression is "a way of thinking about its *Bedeutung*." For them, lack of *Bedeutung* necessarily involves a corresponding lack of *Sinn*. For a discussion, see Dummett 1993a: ch. 7.

Dummett himself criticizes and qualifies the Fregean theory in various ways in *Frege: Philosophy of Language* (e.g. Frege's explanation of how expressions can differ in tone is rejected in chapter 5; Frege's assimilation of sentences to complex proper names and the associated claim that truth-values are a kind of object is questioned in chapter 6; and his claim that expressions other than proper names, such as predicates and quantifiers, also have *Bedeutungen*, is qualified in important ways in chapter 7). However, the suggestion of his that has excited the greatest interest among contemporary philosophers is that, in many important cases, debates in metaphysics between realists and their opponents may and perhaps must be cast as debates in the theory of meaning, between rival accounts of the nature of our grasp of sentences' *Sinne*, or truth-conditions.

Realism

The remarks germane to this suggestion are scattered throughout Dummett's writings, but the key texts are *Truth and Other Enigmas* (1978) (Preface and essays 1, 10, 14, and 21), *The Seas of Language* (1993b) (essays 1–7, 11, and 20), and *The Logical Basis of Metaphysics* (1991c).

This approach to metaphysical questions is indicative of Dummett's view that the philosophy of language – the theory of meaning – has a *foundational* role to play within philosophy. Indeed, Dummett sees this view about the priority of philosophy of language as the defining characteristic of analytic philosophy:

> What distinguishes analytical philosophy, in its diverse manifestations, from other schools is the belief, first, that a philosophical account of thought can be attained through a philosophical account of language, and, secondly, that a comprehensive account can only be so attained. (1993a: 4)

Dummett believes that one of the reasons why philosophical speculation about metaphysical issues has made little progress over the centuries is that the opposing positions in various metaphysical disputes have only been explained in pictorial, or metaphorical terms:

> Even to attempt to evaluate the direct metaphysical arguments, we have to treat the opposing theses as though their content were quite clear and it were solely a matter of deciding which is true; whereas . . . the principal difficulty is that, while one or another of the competing pictures may appear compelling, we have no way to explain in non-pictorial terms what accepting it amounts to. (1991c: 12)

Dummett's approach is intended to remedy this: the metaphysical disputes are recast as disputes about "the correct model of meaning for statements of the disputed class," thus giving the debates some non-metaphorical content, and enabling the disputes to be resolved within the theory of meaning.

What does it mean to say that the truth-conditions of a range of sentences are "realist"? In short, Dummett's answer is as follows: to say that a range of sentences have realist truth-conditions is to say that those truth-conditions are potentially *verification-transcendent*. To say that a truth-condition is potentially verification-transcendent is to say that we may be incapable, even in principle, of determining whether or not it obtains. Thus, consider discourse about the past: intuitively, the sentence "James II suffered a migraine in 1665, on the afternoon of his 32nd birthday" has a truth-condition – James's suffering a migraine on the afternoon in question – and we can say that this condition either obtained or it did not, even though we may have no way, even in principle (because all the evidence appears to have vanished and time-travel is impossible) of determining which of these was the case (cf. ANSCOMBE). Thus, "James II had a migraine on the afternoon of his 32nd birthday" has a truth-condition, and we may be incapable of determining, even in principle, whether that condition obtained or not: it is potentially verification-transcendent. Likewise, consider arithmetical discourse. Goldbach's conjecture, that every even number greater than

two is the sum of two primes, has a potentially verification-transcendent truth-condition: it has a determinate truth-value even though we are incapable of determining what this truth-value is, since we have no guarantee either that a proof of the conjecture will be constructed or that a counterexample – an even number which is not the sum of two primes – will be found.

Thus, sentences about the past and about arithmetic have potentially verification-transcendent truth-conditions: in this sense, Dummett will claim, their truth-conditions are realist. Now, why is the claim that the sentences of a discourse are potentially verification-transcendent a way of cashing out realism about the subject matter of that discourse? In order to see this we have to recall – from the first section – that the *Sinn*, or sense, of a sentence is given by its truth-conditions, and that understanding a sentence consists in grasping its sense. Thus, understanding a sentence consists in grasping its truth-conditions. Any thesis about the truth-conditions of a set of sentences is *inter alia* a thesis about what our understanding of those sentences consists in. In a slogan, a *theory of meaning is also a theory of understanding*. Thus, someone who accepts that the truth-conditions of a region of discourse are potentially verification-transcendent also accepts that our understanding of the sentences of that discourse consists in our grasp of potentially verification-transcendent truth-conditions. And now the connection with realism about that discourse is relatively easy to see. As Crispin Wright, another important British philosopher who has done more than anyone to further Dummett's agenda, puts it:

> To conceive that our understanding of statements in a certain discourse is fixed . . . by assigning them conditions of potentially [verification]-transcendent truth is to grant that, if the world co-operates, the truth or falsity of any such statement may be settled beyond our ken. So we are forced to recognise a distinction between the kind of state of affairs which makes such a statement acceptable, in the light of whatever standards inform our practice of the discourse to which it belongs, and what makes it actually true. The truth of such a statement is bestowed on it independently of any standard we do or can apply; acceptability by our standards is, for such statements, at best merely congruent with truth. Realism in Dummett's sense is thus one way of laying the essential semantic groundwork for the idea that our thought aspires to reflect a reality whose character is entirely independent of us and our cognitive operations. (1992: 4)

In a large class of cases, we can thus conceive of the metaphysical debate between realists and their opponents – anti-realists – in a particular region of discourse D as concerning whether the sentences of D can plausibly be viewed as possessing potentially verification-transcendent truth-conditions.

Realism: the sentences of D have truth-conditions and these truth-conditions are potentially verification-transcendent.

Anti-Realism: the sentences of D have truth-conditions but those truth-conditions are not potentially verification-transcendent.

Note 1: The above cannot be used to characterize all types of debate between realists and their opponents. One characteristic form of opposition to realism about a region of discourse D is the denial (by non-cognitivists, expressivists, or non-factualists) that the sentences in question *have* truth-conditions, whether potentially verification-

transcendent or not. Dummett has admitted that this style of debate between realism and its opponents is prior to that which turns on the possibility of verification-transcendent truth (1993b: 467). In addition, there are also some serious questions about whether the above characterization can adequately capture what is at issue in *ontological* disputes between realists and their (e.g. nominalist) opponents. (For illuminating discussion, see Hale 1997, esp. sect. 3. See also Michael Devitt 1993.)

Note 2: It is far from clear what it means to say that the truth of a sentence is not potentially verification-transcendent. Does this mean that the sentence can be verified by us as we actually are? By someone, somewhere, as they actually are? By someone, somewhere, given some suitable idealization of their present cognitive powers? And what is permissible as a "suitable idealization"? And how can the notion of effective decidability (see below) be extended from the mathematical case to the empirical domain? These questions must be answered if anti-realism is to have any determinate content. (See Wright 1986: 32.)

Note 3: This characterization of realism makes essential use of Frege's idea that understanding a sentence consists in grasping its truth-conditions. It is plausible that Frege himself was a realist in the sense thus characterized. He writes: "A thinker does not create [thoughts] but must take them as they are. They can be true without being grasped by a thinker" (1967: 30).

If a thought – the sense of a sentence – can be determinately true or false even though that thought is not even grasped by a thinker, then the sentence in question can be true or false even though thinkers are incapable, even in principle, of determining its truth-value.

Dummett now suggests that Frege's realism is seriously challenged when we add to his theory of meaning the insights about linguistic understanding to be found in Ludwig Wittgenstein's *Philosophical Investigations*.

Frege and Wittgenstein on the objectivity of sense

The interpretation of Wittgenstein's views on meaning and understanding is a complex and subtle matter: here we can give but the briefest sketch of one of the facets of Wittgenstein's position that Dummett relies on in challenging our intuitively realist picture of the world. (For a more comprehensive treatment, see Miller 1998: chs 5 and 6.) In fact, the Wittgensteinian view about linguistic understanding that features in Dummett's challenge to realism is a development of another insight of Frege's: that sense is, in a way to be explained, objective. According to Dummett, if Frege had followed this insight through to its logical conclusion, he would have seen that it challenges seriously his commitment to realism (1993b: essay 2, §6).

Recall that the sense of a sentence is a thought, and that according to Frege thoughts are in some sense objective, as opposed to subjective or psychological. This is an extremely important part of Frege's position. Indeed, in the introduction to *The Foundations of Arithmetic*, he states the following as the first of his three "fundamental principles": "Always to separate sharply the psychological from the logical, the subjective from the objective" (1953: x).

This applies not only to the senses of sentences, but to the senses of expressions generally. But what exactly does it mean to say that sense is objective and not subjec-

tive? One thing that it means is that grasping a sense – understanding an expression – is not a matter of associating that expression with some subjective item like a *mental image*, *picture,* or *idea*. Frege is quite explicit about the need to distinguish senses, which are objective, from ideas, which are subjective:

> The reference [*Bedeutung*] and sense of a sign are to be distinguished from the associated idea . . . The reference of a proper name is the object itself which we designate by using it; the idea which we have in that case is wholly subjective; in between lies the sense, which is indeed no longer subjective like the idea, but is yet not the object itself. (Frege 1960: 60–1)

The view that understanding an expression consists in the possession of some associated idea or image is one that has a long list of adherents in the history of philosophy. In distinguishing the sense of an expression from any associated idea, Frege was directly attacking this tradition. The classic example of this view of sense can be found in Book III of John Locke's *An Essay Concerning Human Understanding*.

Some creatures who utter, for example, the word "cube" *understand* that word, and some don't. A parrot can say the word, but unlike a normal human speaker of English, the parrot possesses no understanding of what is said. In Fregean terminology, the human speaker grasps the sense of "cube," whereas the parrot does not. But what does this difference consist in? Locke's suggestion is that the word "cube" is, in the case of the competent human speaker, associated with an *idea of a cube* in that speaker's mind, while in the case of the parrot there is no such idea and so no such association. Locke is thus led to the view that understanding an expression consists in associating it with some idea: "Words, in their primary or immediate signification, stand for nothing but the ideas in the mind of him that uses them" (1975: III, ii, 2).

Locke takes ideas to be mental images or pictures: an idea of a cube is taken to be a mental image or inner picture of a cube. This is clear from the way Locke speaks throughout the *Essay*. For example, in his account of memory the talk of ideas is explicitly cashed out in terms of picturing and imagery:

> The ideas, as well as children of our youth often die before us. And our minds represent to us those tombs to which we are approaching; where though the brass and marble remain, yet the inscriptions are effaced by time, and the imagery moulders away. The pictures drawn in our minds are laid in fading colours. (1975: II, x, 5.1)

We could thus sum up Locke's view of sense as follows (where the sense of "cube" determines that it refers to, precisely, *cubes*): a speaker grasps the sense of "cube" if and only if he is disposed to have a mental image of a cube whenever he hears or utters the word. Why does Frege object to this account of sense? Locke's account leads to a tension between the *public* nature of meaningful language, and the *private* nature of ideas and mental images. On the one hand, language is public in that different speakers can attach the *same* sense to their words, and one speaker can *know* what another speaker means by his words. Different speakers can *communicate* with each other in virtue of the common senses that they have attached to their words. On

the other hand, ideas are private. As Locke himself puts it, a man's ideas are "all within his own breast, invisible, and hidden from others, nor can of themselves be made to appear." Also, my ideas, my "internal conceptions," are visible only to my consciousness, and likewise your ideas, your "internal conceptions," are visible only to your consciousness. But we are attempting to give an account of sense, an account that should help explain how we are able to communicate with each other via the use of language; and how can a theory which construes grasp of sense in terms of the possession of private inner items help explain our ability to use language in successful public communication?

Dummett sees Frege's anti-Lockean argument for the objectivity of sense as vitiated by an erroneous construal of mental images as necessarily private, but nevertheless agrees with its upshot (1973: 157–9). But, Dummett suggests, in order to allow for the objectivity of sense, we need to go further than merely denying that grasp of sense is subjective in the manner just outlined: we need, in addition, to construe grasp of sense in terms of *use*.

> Frege's thesis that sense is objective is . . . implicitly an anticipation (in respect of that aspect of meaning which constitutes sense) of Wittgenstein's doctrine that meaning is use . . . yet Frege never drew the consequences of this for the form which the sense of a word may take. (1993b: 91)

In order to allow for the objectivity and communicability of the sense of an expression, grasp of its sense has to be construed in terms of possession of an ability to use it in certain public and observable circumstances. It follows that if speakers possess a piece of knowledge which is constitutive of linguistic understanding, then that knowledge should be *manifested* in speakers' use of the language, that is, in their exercise of the practical abilities that constitute linguistic understanding.

We'll now see how Dummett attempts to challenge realism, by incorporating the Wittgensteinian insight about understanding within the Fregean theory of meaning.

Dummett's challenges to realism

According to Dummett, the debate between realism and anti-realism about a region of discourse is a debate about the nature of the truth-conditions possessed by the sentences of that discourse. Any account of the truth-conditions of a range of sentences will be unacceptable if it cannot cohere with a plausible account of what our *understanding* of those sentences consists in. Dummett's strategy is to argue that the account of linguistic understanding which realism leads to faces serious problems. The metaphysical debate concerning the plausibility of realism boils down to a debate within the philosophy of language.

Why, then, does Dummett think that there are problems with the realistic construal of linguistic understanding as grasp of potentially verification-transcendent truth-conditions? There are two main challenges: the *acquisition* challenge, and the *manifestation* challenge. (For the canonical statement of the former, see 1978: essay 1; for the latter, see essay 14; for a state-of-the-art exposition of both, and of other challenges to realism as conceived by Dummett, see the Introduction to Wright 1986).

Dummett's acquisition challenge

Suppose that we are considering some region of discourse D, the sentences of which we intuitively understand. Suppose, for reductio, that the sentences of D have potentially verification-transcendent truth-conditions. Thus,

1 We understand the sentences of D.
2 The sentences of D have verification-transcendent truth-conditions.

Now, from (1) together with the Fregean thesis that to understand a sentence is to grasp its sense or know its truth-conditions, we have

3 We grasp the senses of the sentences of D: i.e. we know their truth-conditions.

We now add the apparently reasonable constraint on ascriptions of knowledge:

4 If a piece of knowledge is ascribed to a speaker, then it must be at least in principle possible for that speaker to have *acquired* that knowledge.

So

5 It must be at least in principle possible for us to have acquired knowledge of the verification-transcendent truth-conditions of D.

But

6 There is no plausible story to be told about how we could have acquired knowledge of verification-transcendent truth-conditions.

So, by reductio, we reject (2) to get:

7 The sentences of D do not have verification-transcendent truth-conditions, so realism about the subject matter of D must be rejected.

The crucial premise here is obviously (6). Wright puts the point as follows:

> How are we supposed to be able to *form* any understanding of what it is for a particular statement to be true if the kind of state of affairs which it would take to make it true is conceived, *ex hypothesi*, as something beyond our experience, something which we cannot confirm and which is insulated from any distinctive impact on our consciousness? (1986: 13)

However, as Wright notes, this argument is at best inconclusive. It really only presents the realist with a challenge:

> In order to be more than a challenge, [it] would need the backing of a proven theory of concept-formation of a broadly empiricist sort. [And] the traditional theories of that sort have long been recognized to be inadequate. (1986: 15)

The challenge to the realist is thus: give some plausible account of how the knowledge of verification-transcendent truth-conditions which you impute to speakers could have been acquired. Whether or not this challenge can be met by the realist is very much an open question, in the absence of a proven theory of concept acquisition.

Dummett's manifestation argument

Suppose that we are considering region of discourse D as before. Then:

1 We understand the sentences of D.

Suppose, for reductio, that

2 The sentences of D have verification-transcendent truth-conditions.

From (1) and the Fregean thesis that to understand a sentence is to grasp its sense or know its truth-conditions, we have:

3 We grasp the senses of the sentences of D; that is, we know their truth-conditions.

We then add the following premise, which stems from the Wittgensteinian insight that understanding does not consist in the possession of an inner state, but rather in the possession of some practical ability (see the section "Frege and Wittgenstein on the Objectivity of Sense," above):

4 If speakers possess a piece of knowledge which is constitutive of linguistic understanding, then that knowledge should be *manifested* in speakers' use of the language, that is, in their exercise of the practical abilities that constitute linguistic understanding.

It now follows from (1), (2) and (3) that:

5 Our knowledge of the verification-transcendent truth-conditions of the sentences of D should be manifested in our use of those sentences, that is, in our exercise of the practical abilities which constitute our understanding of D.

Since

6 Such knowledge is never manifested in the exercise of the practical abilities which constitute our understanding of D,

it follows that

7 We do not possess knowledge of the truth-conditions of D.

(7) and (3) together give us a contradiction, whence, by reductio, we reject (2) to obtain:

8 The sentences of D do not have verification-transcendent truth-conditions, so realism about the subject matter of D must be rejected.

The basic point is that, so far as an account of speakers' understanding goes, the ascription of knowledge of verification-transcendent truth-conditions is simply *redundant*: there is no good reason for ascribing it. Consider one of the sentences we considered earlier as candidates for possessing verification-transcendent truth-conditions, "James II had a migraine on the afternoon of his 32nd birthday" or "Every even number greater than two is the sum of two primes." The realist account views our understanding of these sentences as consisting in our knowledge of a potentially verification-transcendent truth-condition. But, in Wright's words:

> How can that account be viewed as a description of any *practical* ability of use? No doubt someone who understands such a statement can be expected to have many relevant practical abilities. He will be able to appraise evidence for or against it, should any be available, or to recognize that no information in his possession bears on it. He will be able to recognize at least some of its logical consequences, and to identify beliefs from which commitment to it would follow. And he will, presumably, show himself sensitive to conditions under which it is appropriate to ascribe propositional attitudes embedding the statement to himself and to others, and sensitive to the explanatory significance of such ascriptions. In short: in these and perhaps other important respects, he will show himself competent to use the sentence. But the headings under which his practical abilities fall so far involve no mention of evidence-transcendent truth-conditions. (Wright 1986: 17)

This establishes (6), and the conclusion follows swiftly. A detailed assessment of the plausibility of this argument is impossible here: but we should note that premise (4) depends upon an interpretation of Wittgenstein's work on rule-following and understanding (see WITTGENSTEIN), and that this is an extremely controversial matter (see Miller 1998: chs 5 and 6). In particular, one issue that needs to be addressed is whether the interpretation of premise (4) required by Dummett for the anti-realist argument is left intact by John McDowell's interpretation of Wittgenstein, according to which understanding can harmlessly be construed as a state of mind (see McDowell 1998b: essays 11–14). On Dummett's anti-realist arguments generally, see McDowell 1998a: essays 1, 4, 5, 14, 15, 16. For an excellent survey of possible realist responses to both the acquisition and manifestation challenges, see Hale 1997.

Anti-realism

(1) A sentence is said to be *effectively decidable* if there is some procedure which we can in principle apply and which will guarantee an answer to the question whether or not the sentence is true. Thus, "2 + 2 = 4" and "The Queen had cornflakes for breakfast yesterday" are both effectively decidable: we can carry out an elementary arithmetical calculation in the first case, and we can gather the obvious sorts of evidence in the second case, in order to determine the truth-values of the respective sentences. But "James II had a migraine on the afternoon of his 32nd birthday" and "Every even number greater than two is the sum of two primes" are *not known to be decidable*: in neither case do we know a procedure which we can apply to determine whether or not they are true. Now intuitively, we think that even though these sentences are not known to be decidable, we can nevertheless still assert that they are either true or false: "Every even number greater than two is the sum of two primes" has a determinate truth-value, it's just that we cannot work out what this truth-value is. In other words, even though the sentence is not known to be decidable, we still think that the *principle of bivalence*, that every (non-vague) sentence is determinately either true or false, applies to it. Now this is an idea that is put under pressure by the conclusion of the anti-realist arguments of Dummett's we have been considering. If truth is not verification-transcendent, it is *epistemically constrained*. One way to spell out what it means to say that truth is epistemically constrained is to say that it must be construed in terms of some notion like *correct* or *warranted assertability*: to say that a sentence is true is to say that there is a warrant to assert it, or that it possesses some other prop-

erty that is constructed out of warranted assertability. (This is greatly oversimplified: for more detail, see Wright 1986: §v and 1992: ch. 2, where he suggests that for certain discourses, truth may be modelled on "superassertibility." For another attempt to construe truth as essentially epistemically constrained, see Putnam 1981 (see PUTNAM). See also Tennant 1987, 1997.) Now given that truth is thus epistemically constrained, what can we say about "Every even number greater than two is the sum of two primes"? We do not have a warrant to assert this – since no one has yet been able to construct a mathematical proof of it – nor do we have a warrant to assert its negation – since no one has yet produced a counterexample to it, or established that such a counterexample must exist. Given this, and given that truth is to be to construed in terms of warranted assertability, we cannot assert that the sentence "Every even number is the sum of two primes" is either true or false. That is to say, *we cannot assert a priori the principle of bivalence for sentences that are not known to be decidable*: we cannot assert, a priori, that they are either true or false.

(2) Note that we have here characterized realism as the view that truth is not essentially epistemically constrained, and derived the realist's attitude to the principle of bivalence as a consequence. Dummett himself prefers to characterize realism *directly* in terms of adherence to the unrestricted principle of bivalence (see e.g. 1993b: 230), so that *any* denial of that principle must be seen as inclining one in the direction of anti-realism. But this seems to be a mistake. There are many reasons why the principle of bivalence might fail for a particular region of discourse: because the relevant sentences contain empty names, have false presuppositions, contain vague predicates applied to borderline cases; none of these seem to concern the issue of realism versus anti-realism. So the rejection of bivalence is a symptom of anti-realism, which may or may not signal the rejection of realism depending on whether or not the failure of the principle of bivalence stems from the rejection of truth as essentially epistemically unconstrained. So it is better to characterize realism directly in terms of epistemic constraints on truth, and view issues about bivalence as having only secondary, derivative significance. For an excellent discussion of this point, see Rosen 1995.

(3) Note that if we characterize meaning in terms of an epistemically constrained notion of truth – perhaps in terms of conditions of warranted assertibility – we thereby avoid the problems raised by the manifestation challenge for the realist conception of linguistic understanding. Because the conditions whose grasp constitutes understanding are conditions which, by their very nature, are in principle capable of being recognized whenever they obtain, we can identify grasp of a sentence with a practical ability. This is the ability to discriminate between those recognizable circumstances in which the sentence is true and those which it is not. So that the manifestation challenge has a simple answer when directed at the anti-realist conception of understanding. Likewise for Dummett's acquisition challenge.

(4) Dummett's anti-realist claims that we cannot assert a priori the principle of bivalence, at least as applied to sentences that are not known to be decidable. Now the principle of bivalence – that every (non-vague) sentence is determinately either true or false – is closely associated with the principle of classical logic known as *the law of*

excluded middle: $\vdash P \vee \sim P$. Refusing to assert a priori the principle of bivalence, as the anti-realist proposes, thus appears to threaten the law of excluded middle, and the classical system of logic which is founded upon it. There is much debate among anti-realists about whether anti-realism implies *revisionism* about classical logic: Dummett has argued that anti-realism implies that classical logic must be given up in favor of some form of *intuitionistic* logic which does not have the law of excluded middle as a theorem. (The issues here are complex. For Dummett's own examination of intuitionism see his *Elements of Intuitionism*; for discussion of the alleged revisionary aspects of Dummettian anti-realism, see Wright 1986, and 1992, ch. 2).

(5) It is important to be clear that although the anti-realist claims that we cannot assert that sentences not known to be decidable are either true or false, he is *not* claiming that we *can* assert that they are *neither* true nor false. Dummett is explicit that although the anti-realist does not wish to assert a priori the principle of bivalence, he does not reject the principle of *tertium non datur*, that there is no third truth-value ("neither true nor false") standing between truth and falsity. This might seem puzzling. Suppose that the principle of bivalence corresponds to the law of excluded middle: $\vdash P \vee \sim P$, and that the principle of *tertium non datur* corresponds to $\vdash \sim\sim (P \vee \sim P)$ (is not the case that neither P nor not-P). Since it is a logically valid sequent that $\sim\sim P \vdash P$, doesn't it follow that rejecting $\vdash P \vee \sim P$ entails the rejection of $\vdash \sim\sim (P \vee \sim P)$, that rejection of the principle of bivalence entails rejection of the principle of *tertium non datur*? The crucial point is that the sequent $\sim\sim P \vdash P$ that licenses this entailment is valid in classical logic *but not valid in intuitionistic logic*. Rejection of bivalence entails rejection of *tertium non datur* only given classical logic; but the Dummett-style anti-realist *rejects* classical logic, and so can reject bivalence whilst holding on to *tertium non datur*. (See Dummett 1978).

(6) Finally, note that the anti-realist attitude to sentences which are not known to be decidable is completely different from the logical positivist attitude to sentences which are not in principle verifiable. Whereas a logical positivist such as A. J. Ayer in *Language, Truth, and Logic* claims that such sentences – because they are in principle unverifiable – are literally meaningless, Dummett's anti-realist claim is that sentences that are not known to be decidable *are* meaningful but their meanings have to be construed in terms of an epistemically constrained notion of truth (see AYER).

Limitations and prospects

According to Dummett, to oppose realism, to espouse anti-realism, is to deny that truth is potentially verification-transcendent and to argue that truth must be viewed as epistemically constrained. But is this the best way of cashing out the metaphysical debates between realists and their opponents? We have already mentioned some of the limitations of this approach in the section 'Realism." But even waiving these problems, there are others. For example, although Dummett's way of characterizing the metaphysical debate seems to be appropriate in some cases (e.g. mathematics, statements about the past, statements about the external world) there are other cases where it simply seems besides the point. Consider discourse about, for instance, morals or comedy. It seems

that in these cases a moral realist would not have to claim that the truth-conditions of the relevant sentences are potentially verification-transcendent and that both the moral realist and the moral anti-realist can *agree* that statements about comedy or moral value do not have verification-transcendent truth-conditions. As Wright puts it:

> There are, no doubt, kinds of moral realism [or realism about comedy] which do have the consequence that moral [or comic] reality may transcend all possibility of detection. But it is surely not essential to any view worth regarding as realist about morals [or comedy] that it incorporate a commitment to that idea. (1992: 9)

Intuitively, a sensible version of realism about "That remark was funny" or "That deed was wrong" does not have to view facts about funniness or wrongness as potentially verification-transcendent. So although construing realist truth-conditions as verification-transcendent truth-conditions may be useful for characterizing realism about *some* areas of discourse, there are other areas for which this is not a useful characterization. The upshot of this is that we need *other* ways of fleshing out the notion of a realist truth-condition. The most important recent work on the philosophical agenda initiated by Dummett has consisted of attempts to do just this.

In *Truth and Objectivity*, Crispin Wright argues that non-cognitivism – the denial that the sentences of a discourse are truth-apt or even possess truth-conditions – does not provide a useful way of formulating opposition to realism. The debate between realism and anti-realism about a discourse takes place only *after* it has been granted that the sentences of that discourse are truth-apt. There are two main parts to Wright's sketch of the shape of the debates. First, he develops a version of *minimalism about truth-aptness*, according to which all of the discourses, including morals, comedy, the external world, mathematics, the past, and so on, do turn out to be truth-apt. Wright's approach is thus superior to Dummett's, insofar as he is not content simply to ignore the debate between cognitivism and non-cognitivism about a region of discourse. Second, he develops a *number* of ways of characterizing realism and anti-realism about discourses whose truth-aptness has already been granted – that is, a number of different ways in which truth-conditions can be more or less realist. It turns out that viewing the sentences of a discourse as having potentially verification-transcendent truth-conditions is only one of a number of ways of characterizing realism. Wright's approach is thus superior to Dummett's insofar as it does not involve saddling the moral realist with claims about potential verification-transcendence that any sensible moral realist would baulk at. Both parts of Wright's program have been widely discussed in the literature, and are the point of departure for philosophers wishing to follow up the debate started by Dummett. For more, see Hale 1997 and Miller 1998: ch. 9.

Other work

In the above, I have touched briefly only on those aspects of Dummett's work that I take to be his most important contribution to analytic philosophy. There are many other aspects which lack of space has prevented me from discussing, and I can mention only a few of these here. *Frege: Philosophy of Mathematics*, is Dummett's study of Frege's Platonist and logicist views on arithmetic. Much of the book is critical of the attempts

of Wright and Hale to develop a "neo-Fregean" view of arithmetic in, respectively, *Frege's Conception of Numbers as Objects*, and *Abstract Objects*. For some neo-Fregean responses to Dummett's criticisms, see Hale 1994 and Wright 1994. Dummett's writings on the philosophy of mathematics cannot easily be disentangled from his writings on the philosophy of language, but key papers are 1978: essays 11 and 12 and 1993b: essay 18. For Dummett's introductory survey of the area, see his 1998 paper "Philosophy of Mathematics." Dummett has also done important work on vagueness (1978: essay 15) and causation (1978: essays 18 and 19, 1993b: essay 15).

Bibliography

Works by Dummett

1973: *Frege: Philosophy of Language*, London: Duckworth.
1977: *Elements of Intuitionism*, Oxford: Clarendon Press.
1978: *Truth and Other Enigmas*, London: Duckworth.
1981: *The Interpretation of Frege's Philosophy*, London: Duckworth.
1991a: *Frege and Other Philosophers*, Oxford: Clarendon Press.
1991b: *Frege: Philosophy of Mathematics*, Cambridge, MA: Harvard University Press.
1991c: *The Logical Basis of Metaphysics*, Cambridge, MA: Harvard University Press.
1993a: *Origins of Analytical Philosophy*, Cambridge, MA: Harvard University Press.
1993b: *The Seas of Language*, Oxford: Clarendon Press.
1998: "Philosophy of Mathematics," in *Philosophy* 2, ed. A. Grayling, Oxford: Oxford University Press.

Works by other authors

Ayer, A. J. (1946) *Language, Truth, and Logic*, New York: Dover Press.
Devitt, M. (1993) *Realism and Truth*, Princeton, NJ: Princeton University Press.
Evans, G. (1982) *The Varieties of Reference*, Oxford: Oxford University Press.
Frege, G. (1953) *The Foundations of Arithmetic*, Evanston, IL: Northwestern University Press.
——(1960) "On Sense and Meaning," in *Translations from the Philosophical Works of Gottlob Frege*, ed. P. Geach and M. Black, Oxford: Oxford University Press, pp. 56–78.
——(1967) "The Thought," in *Philosophical Logic*, ed. P. Strawson, Oxford: Oxford University Press.
Hale, B. (1987) *Abstract Objects*, Oxford: Blackwell Publishers.
——(1994) "Dummett's Critique of Wright's Attempt to Resuscitate Frege," *Philosophia Mathematica* 3/2, pp. 122–47.
——(1997) "Realism and its Oppositions," in *The Blackwell Companion to the Philosophy of Language*, ed. B. Hale and C. Wright, Oxford: Blackwell Publishers, pp. 271–308.
Locke, J. (1975) *An Essay Concerning Human Understanding*, ed. P. Nidditch, Oxford: Oxford University Press.
McDowell, J. (1998a) *Meaning, Knowledge and Reality*, Cambridge, MA: Harvard University Press.
——(1998b) *Mind, Value, and Reality*, Cambridge, MA: Harvard University Press.
Miller, A. (1998) *Philosophy of Language*, London: UCL Press.
Putnam, H. (1981) *Realism, Truth, and History*, Cambridge: Cambridge University Press.
Rosen, G. (1995) "The Shoals of Language," *Mind* 104, pp. 599–609.
Tennant, N. (1987) *Anti-Realism and Logic*, Oxford: Clarendon Press.
——(1997) *The Taming of the True*, Oxford: Clarendon Press.

Wittgenstein, L. (1974) *Philosophical Investigations*, Oxford: Blackwell Publishers.
Wright, C. (1983) *Frege's Conception of Numbers as Objects*, Aberdeen: Aberdeen University Press.
——(1986) *Realism, Meaning, and Truth*, Oxford: Blackwell Publishers.
——(1992) *Truth and Objectivity*, Cambridge, MA: Harvard University Press.
——(1994) "Critical Notice of Dummett's *Frege: Philosophy of Mathematics*," *Philosophical Books* 35, pp. 89–102.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

32

Hilary Putnam (1926–)

JOHN HEIL

The Philosophical Lexicon contains the following entry for "hilary":

> **hilary**, n. (*from* hilary term) A very brief but significant period in the intellectual career of a distinguished philosopher. "Oh, that's what I thought three or four hilaries ago." (Dennett 1987: 11)

The entry makes reference to Hilary Putnam's penchant for changing his views, even completely reversing himself on central themes.

What are we to make of this inconstancy? Emerson, in a famous but widely misquoted passage wrote:

> A foolish consistency is the hobgoblin of little minds, adored by little statesmen and philosophers and divines. With consistency a great soul has simply nothing to do. (Emerson 1940: 152)

Philosophers, especially technically adroit philosophers, often adopt a position then stick with it no matter what. If you are clever, you can find ways around almost any objection. Defending a cherished thesis can be like defending an old friend against a charge of dishonesty: your own honor as well as your friend's is at stake. In philosophy, however, candor trumps constancy. It is to Putnam's credit that he has been willing to allow his views to evolve as they will even when this leads him in surprising directions.

Commendable as it is, this kind of intellectual forthrightness puts pressure on anyone setting out to summarize Putnam's views. Not only has Putnam written on a wide variety of issues, but his take on those issues has shifted, often dramatically. There is not one Putnam, but many. In what follows I have selected from among the available Putnams those that seem to me to have had the most immediate philosophical impact. Inevitably, I have had to leave out much that is interesting. I shall not discuss any of Putnam's important technical work in the philosophy of mathematics, logic, and the philosophy of science. (His views on these and many other topics can be found in Putnam 1975a, 1975b, and 1983.) I shall focus – selectively – on three domains in which Putnam has had considerable influence over philosophy as it is now practiced: philosophy of language, philosophy of mind, and metaphysics.

Philosophy of language

Suppose you utter the sentence "That's a banana" meaning to indicate a banana. What is it about you that makes it the case that your utterance, accompanied perhaps by a gesture, indicates a banana? One natural response is that the pertinent feature of you that make it the case that your utterance concerns a banana is your own state of mind. When you deliver the utterance, your mind is focused in a certain way (on bananas!), and this mental focusing is what gives your linguistic utterance its significance. I *understand* your utterance as indicating a banana when the utterance triggers in me a comparable state of mind. This state of mind constitutes my *grasp* of your utterance's significance.

Proponents of the traditional *ideational* model of meaning appeal to ideas or mental images. As you utter the sentence, you entertain a banana image. The image secures a connection between your words and the world. Imagery is not essential, however. A meaning might be a kind of definition you carry around inside your head, a recipe that tells you when to apply a particular term. (Let us ignore a regress problem that seems to undercut any such view: if we understand terms only by possessing a definition, what enables us to understand the terms constituting that definition?) Think of images and mental recipes as mechanisms for fixing the extension of terms. (The extension of a term is the set of objects it designates. The English word "water," for instance, designates a kind of stuff. When you use this word, you designate that stuff.)

Approaching the topic from this direction invites us to distinguish sharply between *meaning* and *reference*. The *meaning* of "water" includes only elements that you the speaker can grasp. Competent speakers need know nothing of the chemical composition of water; they may be ignorant of the nature of the stuff designated by the term "water." Even those with an intimate knowledge of the constitution of water are rarely in a position to apply this knowledge to determine when a liquid substance is or is not water. If we eliminate specialized knowledge as a requirement for knowing the meaning of "water," we are left with the idea that the term means a clear, colorless, tasteless liquid found in oceans, ponds, and rain puddles. Your grasp of this meaning is what enables you to use the term "water" correctly. The meaning, understood as a kind of recipe or rule grasped by speakers, is what connects the word "water" to the stuff, water.

Wittgenstein inaugurated a sustained attack on views of this kind in the period between the two world wars (see Wittgenstein 1953). According to Wittgenstein, meaning is determined by social and contextual factors. Your utterance means what it does, not because of a state of mind that lies behind the utterance, but because you produce the utterance as a member of a particular community of language users. In this community words are seamlessly integrated with actions and with interactions with non-linguistic states of affairs. To understand the meaning of a word is to understand how it figures in the practices – linguistic and otherwise – of members of your community (see WITTGENSTEIN). This understanding is ultimately grounded in your capacity to *engage* in the pertinent practices.

Putnam's approach to meaning can be seen as involving an articulation and extension of Wittgenstein's insight. Putnam begins with an attack on the familiar distinction between meaning and reference. We learn to use the term "water" at an early age.

Later we learn that the stuff to which "water" refers is H_2O. In this way we discover empirically what water *is*. This discovery, however, sheds light as well on what our term "water" *means*. In speaking of water we mean to be speaking about stuff like *this* stuff (here we point to some water). We may know nothing of the stuff's hidden nature, but we use "water" to designate the stuff with the nature of *this* stuff, whatever that might be. The stuff in question is, as we now know, H_2O. So water's *being* H_2O is part of the *meaning* of "water."

Twin Earth

To reinforce this point, Putnam invites us to imagine a distant planet that precisely resembles Earth in almost every respect (see Putnam 1975b: ch. 12). Its continents are arranged exactly as our continents are arranged, its inhabitants speak languages precisely resembling languages spoken on Earth. Were you instantaneously transported to this planet, you would notice no differences at all. Inhabitants of the planet who live in a place they call "America" refer to their planet as "Earth," but to avoid confusion let us dub it Twin Earth. Twin Earth differs from Earth in just one respect: the colorless, tasteless, odorless liquid stuff that fills Twin Earth oceans, rivers, ice trays, and fish tanks is not H_2O, but a different substance, XYZ. Although XYZ superficially resembles H_2O, it possesses a very different chemical constitution. In other respects, however, Twin Earth precisely resembles Earth down to the last detail.

Suppose that you utter the sentence, "I'll have a glass of water, please." Your utterance concerns water and, assuming it occurs in appropriate circumstances (you are not rehearsing for a play, for instance, or making a philosophical point), you are issuing a request for a glass of water. Imagine, now, that your twin on Twin Earth produces an exactly resembling utterance. Your twin's utterance does not concern water, nor does your twin request a glass of water. Water is H_2O, and the stuff called "water" on Twin Earth is not H_2O, but XYZ. We might say that your twin's utterance of "water" concerns *twin* water; your twin is requesting a glass of *twin* water.

You and your Twin Earth counterpart may be as alike as you please (leaving aside the fact that the chemical constitution of your respective bodies will be importantly different!); the images running through your mind could precisely resemble the images running through your twin's mind; your feelings could be the same. Yet your utterance and your twin's appear to have different meanings. Given the intrinsic similarities between you and your twin, the meanings of words you utter must be determined by something other than your intrinsic makeup. You and your twin may be entirely ignorant of the chemical constitution of what you each call "water." Indeed, English speakers (and their counterparts on Twin Earth) used the term for generations before anyone was in a position to appreciate that water was a particular sort of chemical compound. Earlier, we regarded this as a good reason to suppose that the meaning of words must be limited to what speakers can individually grasp. But this conception of meaning as strongly distinguished from reference is precisely what Putnam's Twin Earth thought experiment challenges. If we think of the meanings of our terms as what fixes the extension of those terms – where the extension of a term is just the stuff or set of objects designated by the term – then we must give up the idea that meanings are like pictures or recipes we carry around inside our heads and consult when we apply words to objects.

The division of linguistic labor

English speakers use the terms "beech" and "elm" to designate species of tree. If you are like me, however, you would be hard pressed to say how beeches and elms differ and utterly unable to distinguish a beech from an elm in the wild. Does this mean that, for English speakers who lack a capacity to tell beeches and elms apart, the words "beech" and "elm" are synonymous? That seems implausible. Putnam suggests that, when it comes to such *natural kind* terms, we rely on a division of linguistic labor. (Natural kind terms – "gold," "water," "planet," "tiger" – designate stuffs and objects thought to occur naturally, and are distinguished from *artifactual kind* terms: "table," "senator," "dollar bill.") We use "elm" and "beech," for instance, to designate species of tree *that would be so labeled by experts*.

Twin Earth cases and the phenomenon of the division of linguistic labor make it clear that agents who are indiscernible in all relevant intrinsic respects could nevertheless differ in what they mean by their utterances. If this is right, accounts of meaning that focus solely on agents considered in isolation are bound to fail. An adequate account of meaning apparently brings with it a battery of social and contextual elements. Putnam puts it succinctly: "Cut the pie any way you like, 'meanings' just ain't in the *head!*" (1975b: 227). The philosophical impact of this thesis – I shall call it *externalism* – would be hard to overstate. As we shall see, what goes for meaning goes for thought as well. If Putnam is right, the traditional conception of the mind as a spectator on the "external world" must be abandoned, replaced by a conception of the mind as constituting – and constituted by – the world.

This is to get ahead of our story, however. Let us look first at Putnam's articulation and defense of functionalism, a conception of the mind according to which minds comprise systems of relations among elements that resemble the states of a computing machine.

Philosophy of mind

In the 1950s, English-speaking philosophers under the spell of Wittgenstein came to regard philosophical questions as expressions of linguistic befuddlement. We ask, for instance, "What is truth?" and interpret this as a substantive question, one that calls for the investigation of some independently existing reality. Wittgenstein argued that such questions occur to us only when we distance ourselves from linguistic practices that give form to talk of truth (or any other philosophically challenging concept).

Augustine (*Confessions*, XI, xiv) remarked about time: "What then is time? If no one asks of me I know; if I wish to explain to him who asks, I know not." The sentiment (though not Augustine's subsequent treatment of it) is profoundly Wittgensteinian. So long as we *use* language in the pursuit of ordinary human ends, we remain innocent of philosophy. We are moved to philosophical questioning when we step outside the linguistic practices that ground our use of words. We ask "What is time?" and seek an answer in a way that ignores the way "time" and its cognates are actually deployed in our linguistic community. Once we lose our moorings within language, our theorizing is colored by a misapprehension of

the roles of terms that generate familiar philosophical puzzles. This misapprehension is systematic: the same kinds of theory arise over and over in the history of philosophy.

Wittgenstein's positive proposal is deflationary. Philosophical puzzlement requires treatment. When a philosopher re-immerses himself in the linguistic practices and forms of life that give sense to the terms he finds bewildering, the bewilderment ebbs. Philosophical questions are not answered but laid to rest. The temptation to pose such questions, although perfectly natural, requires a kind of therapy (see WITTGENSTEIN). The philosopher who responds to this therapy is no longer impelled to philosophize; the end of philosophy is the end of philosophy.

Wittgenstein's approach to philosophical issues concerning the mind led to the rejection of the traditional idea of minds as mental organs that receive inputs via the senses and yield outputs in the form of utterances and bodily motions. "Mind" is a substantive noun, but talk of minds is not talk of a substance or entity associated with, but somehow distinct from, the body. On the contrary, in regarding you as possessing a mind, I regard you as engaging in intelligent activities, responding to the world in intelligible ways, and so on. Thoughts like these led, in turn, to philosophical behaviorism (see RYLE): possessing a mind is exclusively a matter of behaving, or being disposed to behave, in particular ways (see, e.g., Ryle 1949; for a response, see Putnam 1975b, chs 14, 15, 16). Behaviorists hoped to analyze or translate talk of mental goings-on (feelings, thoughts, intentions) into talk of behavior and behavioral dispositions. To be depressed, for instance, is not to be in a particular kind of inner state, but to mope about, complain, or be disposed to complain, and the like. Behaviorists need not deny that inner states *accompany* bouts of depression, only that these inner states *are* the depression.

One difficulty for the behaviorist program stemmed from the fact that behaviorist analyses of mental concepts typically included reference to other mental concepts. Someone who is depressed, for instance, is disposed to form thoughts of certain sorts, and to acquire (or lose) certain motives and desires. When we attempt to analyze *these* mental concepts behavioristically, we find we must appeal to other mental concepts; analyses of these concepts require reference to still other mental concepts; *and so on*. (As we shall see, the interconnectedness of mental concepts comes to the fore with the development of behaviorism's intellectual successor, functionalism.)

The analytical program of behaviorism was challenged, first by the advent of the mind–brain identity theory (see Place 1956, Smart 1959) and then by functionalism. Mind–brain identity theorists defended the thesis that conscious states were at bottom states of brains. They argued that the kinds of correlation known to hold between subjects' reports of states of consciousness and states of those subjects' brains are best construed as evidence for the identification of states of consciousness with brain states. Imagine that, while shopping, you drop a can of tomato soup on your foot and, as a result, you experience a throbbing pain in your big toe. Neuroscientists tell us that, when you experience a pain of this sort, certain kinds of event occur in your brain. (Let us pretend that pains are associated with the firing of C-fibers in the spinal cord.) Identity theorists argued that the best explanation of the correlation between C-fiber firings and reports of pain was that being in pain just *is* the firing of C-fibers.

Functionalism

Nowadays many scientifically minded theorists regard it as close to obvious that states of mind are brain states, mental events are neurological events. Professional philosophers, however, have by and large resisted this conclusion. This is not because philosophers have a preference for mind–body dualism, but because most philosophers have been convinced by arguments pioneered by Putnam that the mind–brain identity theory suffers a fundamental defect (see Putnam 1975b, chs 18, 19, 20).

Consider the fact that we unhesitatingly ascribe states of mind to creatures other than human beings. Think of being in pain, and suppose for the sake of argument the mind–brain identity theory were correct: pains *are* brain states; your being in pain *is* your being in a particular kind of brain state: your C-fibers are firing. So far, so good. But now consider: can an octopus feel pain? It surely seems so. The neurological makeup of an octopus is very different from the neurological makeup of a human being, however. This seems to imply that octopodes, sporting a different physiology, lack a capacity for pain: if pains are C-fiber firings, and octopodes' pain responses are triggered by different mechanisms (as they surely are), then octopodes do not feel pain! Suppose we encountered intelligent creatures from distant planets who were like us in many ways but whose biology was silicon-based. We might have excellent grounds for regarding such creatures as experiencing pain, yet, if the mind–brain identity theory were true, this would be impossible if such creatures lacked C-fibers (as they almost certainly would). If having a pain is a matter of being in a particular kind of neural state, no creature lacking such states could experience pain.

If human beings, octopodes, and Alpha Centaurians can all experience pain, then it is hard to see how pain could be identified with kinds of brain state found only in human beings (and their near relations). Suppose, however, we think of pain states, not as neurological states, but on the model of *computational* states. Reflect on an ordinary desktop computer. The device's operation is governed by *programs* that it runs. When you elect to print a document you have been working on, your desktop computer runs a simple program that sends signals to a printer, which then prints the document. Now suppose we distinguish between the program your desktop computer is running and a particular physical *implementation* of that program. The machine's running the program is a matter of its going into a sequence of physical states. These states, we might say, *realize* the program. Note, however, that a different machine could run the very same program by going into a sequence of very different kinds of physical state. In the 1950s, computing machines consisted of ungainly arrays of vacuum tubes; modern computers make use of tiny transistors; in the nineteenth century, Charles Babbage constructed a sophisticated computing machine using brass gears and cylinders; and today there is talk of molecular computers. It is possible for all of these devices to run the very same program, to engage in the very same sequence of computations, and so to encompass the very same computational states.

Distinct machines can be in the same computational state, then, even if they are made of very different physical ingredients. All that is required is an isomorphism – a one–one correspondence – between sequences of operations performed by the machines (and sameness of inputs and outputs). You feed into a simple calculator "7," "+," and "5," and the calculator displays "12." The causal chain leading the calculator through

this computation has a certain physical character. When you type this same sequence into your desktop computer or dial it into a Babbage machine, these devices go through vastly different kinds of causal sequence to arrive at the same output: "12." At any rate, the sequences are vastly different considered solely as physical events. They exhibit, however, a common structure, a corresponding set of relations. You might put this by saying that, considered concretely, the events are very different but, considered at a higher level of abstraction, the sequences they embody are the same.

What has any of this to do with the mind? In suggesting that states of mind are computational states, Putnam is not imagining that creatures with minds – human beings, for instance – are "mere robots," creatures whose actions are inflexible and "mindless." The idea, rather, is that states of mind owe their identity, not to their physical makeup, but to their place within a structured system. To avoid misleading associations, I shall speak henceforth, not of computational states, but of functional states. (Are functional states and computational states co-extensive? This is controversial. A computational state can be given a particular sort of formal characterization. If every functional state or process is characterizable in this formal way, then every functional state is a computational state.)

Functional states are picked out by reference to roles they occupy – functions they perform – within a system. An analogy may help. Wayne is a vice-president of the Gargantuan Corporation. What exactly does Wayne's being a vice-president amount to? Wayne is 175 cm tall, balding, and overweight. These intrinsic properties of Wayne seem not to bear on his being a vice-president. Wayne could be "re-orged," and replaced by Becky, a petit brunette, by Oscar, a robot, by Hans, a chimpanzee fluent in sign language, or even by Renée, an immaterial angel. Wayne is a vice-president, not in virtue of his intrinsic properties, but in virtue of relations he bears to others in the organization in which he occupies this office. Anyone (or any*thing*!) bearing these relations would be a vice-president.

Functionalists contend that what goes for vice-presidents goes for states of mind. Being in pain is not a matter of being in a particular kind of neurological state, but a matter of being in a state that bears the right sorts of relation to other components of the system to which it belongs. In your case, a particular neurological state occupies the pain role; in the case of an octopus or an Alpha Centaurian, very different kinds of physical state fill the pain role. A creature is in a state of pain when it is in a state that is typically caused by tissue damage, and that causes certain characteristic beliefs and desires (the belief that this hurts, for instance, and a desire for the pain to stop), and certain characteristic actions (if you've stepped on a tack you will quickly move your foot). This makes functionalism sound like dressed up behaviorism. Functionalism, however, unlike behaviorism, does not require that states of mind be characterizable solely in terms of stimuli and responses. How you respond to pain – your behavior – can depend partly on what you believe and desire. If you are trying to impress a companion with your toughness, you may shrug off a pain that you would react to very differently were you alone.

You may worry that this way of characterizing mental states is ultimately circular. We designate a mental state by noting its relations to other mental states. These, in turn, are characterized by reference to other mental states. Eventually we come back to the original states.

The threat of circularity is warded off by means of a technique introduced in a different context by Frank Ramsey and refined by David Lewis (see Lewis 1972). The issues are technical, but the guiding idea is straightforward. Imagine that you define states of mind by locating them as nodes in a network of nodes, each of which represents a distinctive kind of mental state. The system of nodes is anchored at one end by relations to incoming stimuli, and at the other end by behavioral outputs. Now the pain node will have a certain unique structural relation to other nodes in the system; a feeling of pleasure will have another kind of structural relation; and a belief or desire will exhibit other kinds of structural relation. We can then say that being in pain is a matter of being in a state exhibiting *these* kinds of relation to elements in a system with *this* kind of structure.

Putnam summarizes this line of reasoning by describing states of mind (indeed computational or functional states generally) as *multiply realizable*. This means that the very same state of mind can be realized by many different kinds of physical (or perhaps nonphysical, ectoplasmic or angelic) state. If one state can have many realizers, that state cannot be identified with or reduced to any of its realizers. Thus, although states of mind are possessed by ordinary conscious agents *by virtue of* those agents' being in some physical realizing state, mental states are not reducible to the physical states that realize them – or so functionalists contend.

Despite its immense popularity, functionalism has been widely criticized. Putnam himself has been among its most vocal critics. Even so, it is fair to say that functionalism remains hugely influential, both inside and outside philosophy. Functionalism provides a way of understanding how mentality could be housed in the brains of human beings (and in the nervous systems of other intelligent species). In addition, functionalism leaves room for distinctive *levels* of explanation. We might come to understand the behavior of a computing machine by investigating its physical makeup or by studying its program. In the same way, you might explain my behavior by citing complex processes in my central nervous system or by reference to my beliefs, desires, and intentions. Of course, the physical makeup of a computing machine might be extremely complicated, and the physical makeup of a human being more complicated still. In most cases, a purely physical explanation of the behavior of either would be a practical impossibility. Nevertheless, functionalism provides a way of seeing how we could be warranted in offering "higher-level," functional explanations of the behavior of complex systems, and doing so in a way that does not compete with lower-level, purely physical, explanations.

Functionalism spurned

Recall Putnam's line on meaning: the meaning of your utterances depends, not merely on your intrinsic features, but on relations you bear to your surroundings. When you utter "that's water," for instance, your utterance concerns water (H_2O) only if you stand in an appropriate relation to water. When your Twin Earth counterpart produces an indistinguishable utterance, that counterpart says something different. When you speak of elms and beeches what you mean is determined in part by experts in your linguistic community who are in a position to identify and distinguish elms and beeches. In this regard, meanings are community affairs. You and the expert mean the same

when you speak of elms, even though you lack the expert's knowledge of the distinguishing marks of elms. Your beliefs about elms might be largely false, still you mean by "elm" what others in your linguistic community mean: your talk of elms is talk of *elms*.

So far, this is a thesis about the meaning of utterances. The thesis is easily extended, however, to the contents of our thoughts: what those thoughts concern. Imagine that, on Earth, Debbie is anticipating a cool drink of water on a hot day. Debbie entertains a thought she would express by saying "that's water." At the same time Debbie's counterpart on Twin Earth, Twin Debbie, is entertaining a thought *she* would express by saying "that's water." Debbie's utterance and thought concern water. Twin Debbie's utterance and thought, in contrast, are not about water; water, after all, is H_2O, and Debbie's utterance and thought are not about H_2O; they are about XYZ, *twin* water! Of course, Twin Debbie *calls* twin water "water," but that is another matter. Debbie and Twin Debbie's uses of "water" resemble the use of "burro" by a Spanish speaker and an Italian. In the mouth of a Spanish speaker, "burro" means donkey; uttered by an Italian, "burro" means butter.

Putnam holds that cases like these make it clear that what our thoughts concern, as well as what our words mean, is fixed, not solely by what is inside our heads but by relations we bear to the world around us. "Water" in Debbie's (but not Twin Debbie's) mouth means water in part because Debbie (but not Twin Debbie) stands in an appropriate causal relation to water – H_2O. Similarly, thoughts Debbie (but not Twin Debbie) would express by utterances featuring the word "water" concern water – H_2O – in part because Debbie (but not Twin Debbie) stands in an appropriate causal relation to water. Twin Debbie's "water" utterances and thoughts she would express by means of these utterances concern, not water, but twin water, XYZ. To be sure, there is no relevant *internal* difference between Debbie and Twin Debbie (ignoring the fact that Debbie's constitution includes H_2O, Twin Debbie's, XYZ). The contents of our thoughts, like the meanings of our words, depend on our context, most particularly on causal relations we bear to objects in the world and social relations we bear to others in our linguistic community. These external relations are partly constitutive of the meanings of words and the contents of thoughts.

It is natural to extend the externalist thesis that meanings are not "in the head" to the meanings or contents of states of mind. This seems to imply that states of mind, or at any rate the contents of states of mind, are not in the head! (If you think that the content of a state of mind is essential to it – the belief that snow is white is essentially the belief that snow is white – then an externalism about content straightforwardly yields an externalism about states of mind with content: beliefs, desires, intentions, and the like.) Surely this is ridiculous! Or is it?

Before trying to answer this question, let us reflect on the implications of all this for functionalism. A functional state is a state of a system, a state definable wholly by relations it bears to other states of the system (and to inputs and outputs). Twin Earth cases, however, appear to show that distinct agents (Debbie and Twin Debbie, for instance) could be functionally identical yet differ mentally: one is thinking of water, another of twin water. You might regard functionalism as framing a more or less traditional "internalist" conception of the mind and its contents. If you are attracted to externalism – a broadly contextual account of the

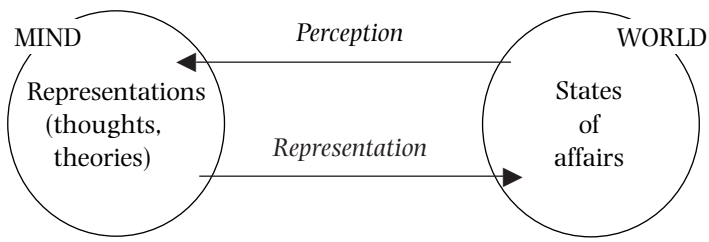
contents of states of mind – you thereby abandon the traditional view of minds as self-contained entities radiating thoughts onto an "external world." The boundary between mind and world becomes blurred: "the mind and the world jointly make up the mind and the world" (Putnam 1981: xi).

If functional states are purely internal states of a system (states definable by their relations to other states of the system), however, and if states of mind are not, then states of mind are not functional states. We are in this way led back to the idea that making sense of the mind and its contents is a matter of seeing intelligent agents in context. The attempt to locate minds inside heads is analogous to an attempt to characterize chess pieces solely by reference to their intrinsic features.

Metaphysics

All this brings us to Putnam's attack on "metaphysical realism," a doctrine, rarely articulated but widely taken for granted, according to which the mind and the world are separated by an epistemological chasm. Minds respond perceptually to, and represent how things stand in, the world. Nevertheless minds are depicted as occupying a standpoint *outside* the world.

Although most of us are by and large committed to the view that minds are physical constituents of the world, it can still seem perfectly natural to *represent* the relation minds bear to the world in a way that situates thoughts "in here" and their worldly objects "out there." We are spectators on the world. Perceiving is a matter of the world affecting the mind (the incoming arrow in the figure), and thinking about the world is a matter of aiming thoughts at the world (the outgoing arrow).



Descartes's mind–body dualism is only one extreme form of metaphysical realism. Indeed, Putnam holds that the metaphysical force of Cartesian dualism lies not in the contention that minds are immaterial entities, but in the idea that minds stand apart from the world on which their thoughts are directed. This picture, the core of metaphysical realism, survives the transition to materialist conceptions of the mind. The result is a kind of internal instability. On the one hand, modern science encourages us to regard minds as material objects alongside other material objects. On the other hand, metaphysical realism depicts the mind as a spectator on the world, seemingly locating minds outside the world they represent. If we insist on situating minds in the world, however, we must abandon metaphysical realism.

One way to get at all this is to consider the implications of meaning externalism. Externalism undermines the conviction that meanings and mental contents are "in the head." What you mean when you utter an English sentence and what your thoughts are directed on, depends in part on your circumstances (and not merely on your constitution or internal organization). The point can be illustrated by means of a simple analogy (borrowed from Wittgenstein). Imagine a picture of a smiling face. Imagine the face appearing in the foreground of a depiction of a child's birthday party. Here the face expresses benevolent happiness. Now imagine the face situated against a background of horrible suffering. In this context, the face expresses evil. Perhaps our thoughts are like this. The same *form* of thought in one context expresses one content, and, in a different context expresses something entirely different. ("Form" here refers to the "shape" or "intrinsic character" of a thought. "Burro" in Spanish encompasses donkeys, in Italian it designates butter. The same linguistic form can have different meanings in different linguistic settings.)

This is one component of Putnam's attack on metaphysical realism. A second component is epistemological. In sharply distinguishing representations of the world – beliefs and theories – and the world those representations concern, we create an unbridgeable chasm. Our representations purport to "match" reality, but we are in no position ever to effect a comparison. At best we can measure representations against other representations. Suppose you believe the ice is thin. You decide to check your belief by examining the ice. You are not measuring your belief against the ice, but measuring your belief against other perceptually induced beliefs about the ice.

If you regard this as a natural and unavoidable feature of the human predicament, you are at least a closet metaphysical realist. One consequence of such a view is that it opens the door to "external world skepticism": what grounds could we have for thinking that our beliefs about the "external world" are true? If we have access only to our own representations, then we could have no assurance that those representations "match" the reality they purport to represent, or even whether there *is* any external reality beyond the representations themselves!

The situation is one Descartes dramatized by imagining an evil demon. The evil demon has the power to make our beliefs about the world false. Descartes's attempt to reconcile metaphysical realism with the conviction that, properly pursued, knowledge of the external world was attainable, involved appeal to a benevolent God: God is such that he would not let us err concerning truths we find indubitable. This gives Descartes a foundation on which to erect an account of knowledge according to which we are entitled to be confident that our beliefs are true provided those beliefs are not based on unproven assumptions. The argument's appeal to a benevolent God, however, strikes most readers as unconvincing.

Note that the skeptical challenge – what entitles us to suppose that our representations of the world match the world? – presupposes metaphysical realism. Skepticism is metaphysical realism seen in the mirror of epistemology (Heil 1998). Metaphysical realism makes the skeptical question inevitable, and the skeptical question makes sense only given the kind of mind–world separation that makes up the core of metaphysical realism. A refutation of realism, then, can be seen as a refutation of "external world" skepticism – and vice versa. This is precisely Putnam's strategy.

Brains in vats

Appeals to a benevolent God aside, let us follow Putnam in updating the skeptical challenge by asking what grounds we have for believing that we are not *brains in vats*. Imagine that you have been kidnapped by an evil scientist, drugged, and your brain removed from your body and kept alive in a vat of nutrients. Nerve endings previously attached to bodily organs are attached now to a super computer. The computer precisely simulates incoming nerve impulses. Nervous stimulation that two days ago would have come from your retina, for instance, now issues from the computer. As far as you can tell, the world is unchanged. Your visual and auditory experiences, even the kinesthetic feedback you receive when you seem to move your body are fed to you by the computer. All this, although undoubtedly fanciful, seems at least physically possible. But, the skeptic insists, if it is *possible*, how could we ever be in a position to know (or even reasonably believe) that we are not brains in vats?

Note, first, that the brain-in-a-vat possibility *is* a possibility only so long as we accept metaphysical realism and its attendant gap between how the world is and how we represent it as being. Does this give us a reason to abandon realism? It hardly seems so. If realism implies that we might be brains in vats, then this is something we shall have to live with. (If you are inclined to dismiss the possibility as idle, ask yourself what grounds you have for dismissing it.)

Here Putnam goes on the offensive. Suppose metaphysical realism implies that we might be brains in vats. If we could establish that it is *not* possible that we are brains in vats, then we will have established that metaphysical realism is false. But how could anyone hope to prove that it is not possible that we are brains in vats? We have, after all, granted that the envisaged envatting of a brain lies within the realm of physical possibility.

Recall Putnam's take on meaning. The meaning of an utterance (or the content of a thought you might express with that utterance) depends on context, most especially on relations speakers and thinkers bear to their surroundings. In the simplest case, your thought of a tree concerns this tree because this tree (and no other) is causally responsible for it. Suppose we generalize this observation. The English words "brain" and "vat" mean what they do in part because members of the English-speaking linguistic community stand in appropriate causal relations to brains and vats. Similarly, your thoughts of brains and vats, thoughts you might express using the terms "brain" and "vat," concern brains and vats because you are, as an English speaker, an agent standing in appropriate causal relations to brains and vats. The causal relations in question are no doubt complex, and it would be difficult to spell them out in detail. But, as the Twin Earth cases seem to show, the presence of an appropriate causal connection is at least a necessary condition for our words and thoughts connecting to the world. Debbie's utterance of "water" designates water, and thoughts she would express using this term concern water, in part because Debbie's is causally related to water. Twin Debbie's utterances of "water" differ in meaning and her corresponding thoughts differ in their content because Twin Debbie stands in comparable causal relations, not to water, but to XYZ, twin water.

Let us allow that "vat" and "brain" mean what they do in part because those of us who deploy these terms stand in certain causal relations to vats and brains. Now

consider an envatted brain, Evan. Consider, in particular, Evan's causal links to the world outside the vat and their bearing on the significance of Evan's "utterances" of "brain" and "vat" (and thoughts Evan might "express" using these terms). Sources of stimulation for these utterances and thoughts are not brains and vats, but electrical events inside the super computer to which Evan is attached. These electrical events – stand-ins for real brains and vats – produce sensory experiences in Evan that precisely resemble the sensory experiences you might have when you encounter a vat or a brain. Evan's situation obliges us to reconstrue Evan's utterances and thoughts, just as we did in the Twin Earth case. When Evan has a thought that he might express by uttering the sentence "That's a brain," we should have to interpret his utterance as meaning something like "That's electrical state s_1." Similarly, when Evan harbors a thought he would express by uttering "That's a vat," we must interpret this utterance as expressing something like what we would express in English as "That's electrical state s_2."

Evan's utterances will need to be systematically reinterpreted. His utterances *formally* resemble English utterances, but they differ significantly in what they *mean*. We can mark this systematic difference by describing Evan as "speaking," not English, but *Vat*-English, just as your twin on Twin Earth speaks Twin English, not English. Of course, just as your twin calls Twin English "English," so Evan calls Vat-English "English." When Evan "says" "I speak English," his "utterance," translated into English (*our* language), means "I speak Vat-English," and this utterance is true.

With all this as background, we are in a position to appreciate Putnam's anti-skeptical argument. Suppose you entertain a thought that you would express by uttering the sentence "I am a brain in a vat." If you are an English speaker, then this sentence is false. Why? If you are an English speaker, you are connected in a normal way with brains, vats, trees, and the like and *not* plugged into a super computer. If you are an English speaker, then, you are not a brain in a vat.

Astute readers will be quick to point out that it is hard to take much comfort from this fact. True, if we grant meaning externalism, we are not brains in vats *if we are English speakers*. But what gives us the right to assume that we *are* English speakers? After all, if we were brains in vats, we would not be English speakers! It looks as though, in order to know that we speak English (and not Vat-English) we should first have to know that we are not brains in vats. We cannot, then, under pain of circularity, appeal to this fact (or alleged fact!) to establish that we are not brains in vats.

Let us think a little harder about this. Consider Evan, and his "utterance" of "I am a brain in a vat." Evan, as *we* know, is envatted and "speaks," not English, but Vat-English. Evan's "utterance" of "I am a brain in a vat" translated from Vat-English into English would mean something like "I am a computer state of type s_n." But this utterance is manifestly false: Evan is a brain in a vat, not a computer state! It appears that you can generalize this point and use it in simple argument to the conclusion that you are not a brain in a vat:

1 If I am a brain in a vat, I express a falsehood in uttering the sentence "I am a brain in a vat."
2 If I am not a brain in a vat, I express a falsehood in uttering the sentence "I am a brain in a vat."
3 I am a brain in a vat or I am not a brain in a vat.

4 In uttering the sentence "I am a brain in a vat," I express a falsehood. (From (1), (2), and (3))
5 In uttering the sentence "I am a brain in a vat," I utter a sentence meaning that I am a brain in a vat.
6 I am not a brain in a vat. (From (4) and (5))

This argument is valid: its premises logically imply the conclusion. But is the argument sound? Does it establish what it purports to establish? Does the argument enable you to prove that you are not a brain in a vat?

Before taking up this question, let us remind ourselves why the issue is important for Putnam. Metaphysical realism, in bifurcating mind and world, implies that it is possible that we are massively deluded: it is possible that we are brains in vats. If we can exclude this skeptical possibility, we shall have thereby established the inadequacy – the falsehood or perhaps incoherence – of metaphysical realism.

Does Putnam's argument work? Suppose that, after studying the argument you conclude that it fails to show that you are not a brain in a vat: *you might be a brain in a vat anyway*. Suppose you express this possibility via the sentence "I am a brain in a vat." As we have seen, your utterance of this sentence is false if you are an English speaker, and it is false if you speak Vat-English. In either case it is false. Assuming (for simplicity) these are the only possibilities, the sentence must be false. Yes, you might reply, but it is vital to know what the false sentence *means*. Premise (5) of the argument tells us that the sentence means that you are a brain in a vat. If you knew this, and knew as well, that the sentence was false, you *would* know that you are not a brain in a vat. But why should we accept premise (5)?

Remember, we are supposing that you are running through the argument in an effort to establish that you are not a brain in a vat. Imagine that you have just run through premise (5). How *could* premise (5) express a falsehood? To see the difficulty in denying (5), pretend that we are eavesdropping on Evan as he runs through the argument (the computer connected to Evan includes a loudspeaker so that we can eavesdrop on Evan's ruminations). When Evan reaches premise (5), he concludes: "In uttering the sentence 'I am a brain in a vat,' I utter a sentence meaning that I am a brain in a vat." Evan is "speaking" Vat-English, so we shall need to translate this utterance into English. When we do so we obtain something like: "In uttering the sentence 'I am a brain in a vat,' I utter a sentence meaning that I am a computer state of type s_n." This utterance is certainly true. More generally, utterances of the form "in uttering the sentence '*P*', I utter a sentence meaning *P*," are bound to be true.

If Putnam is right, then the meanings of these sentences depend on all sorts of causal and contextual factors. We need know nothing of these, however, for them to determine the meaning of what we say and the content of what we think. Debbie's utterances of "water," and thoughts she would express by such utterances, concern water, not because Debbie has figured out that she is causally connected to H_2O (and not XYZ), but because Debbie *is* causally connected to H_2O (and not XYZ).

Imagine, again, that we are eavesdropping on Evan running through Putnam's argument. Pretend that Evan finds the argument convincing, concluding in an excited tone of voice: "I am not a brain in a vat!" Surely, you think, Evan is deluded. As we can plainly see, Evan *is* a brain in a vat. This is too quick, however. Before we can evaluate

Evan's conclusion, we must translate it from Vat-English into English. When we do so, we obtain something like: "I am not a computer state of type s_n!" This, of course, is true; Evan is *not* deluded. We can apply this lesson to our own case: given meaning externalism, we could not be deluded in concluding from Putnam's argument that we are not brains in vats.

Implications for metaphysical realism

Evan is certainly not deluded in the sense of believing a falsehood. He believes that he is not (as we should put it) a computer state, and he is correct. Indeed, most of Evan's beliefs about his actual situation are correct. (The argument is not affected by supposing that Evan has false beliefs; all of us have our share of false beliefs.) Of course Evan cannot appreciate that he is a brain in a vat hooked to a computer programmed by an evil scientist. The thought is not one Evan is in a position to entertain.

Perhaps this is all Putnam needs to show that metaphysical realism is untenable. Metaphysical realism presumes an epistemological gap between what we take to be the case and what is the case. One way to express this is in terms of the possibility that our beliefs about what is the case are massively false. If we accept externalism about the meanings of our words and the contents of our thoughts, if we suppose that what we mean and what our thoughts concern is fixed by our circumstances, we thereby exclude the possibility of massive error. This is what Putnam's argument shows.

If we return to Evan, however, we can see that, although Evan may not be massively deceived, a gap remains between what he takes to be the case and what is the case. The fact, if it is a fact, that Evan is in no position to entertain thoughts concerning what is the case – thoughts about brains and vats – provides scant comfort when we consider our own circumstances. We can still envision a gap, even if we cannot envision what might lie on the far side of the gap. And this evidently leaves realism standing.

Is this unfair? We are imagining that our circumstances might be wildly different from what we take them to be even though our beliefs about those circumstances are, on the whole correct: the deep truth is not thinkable by us. But what sense could be made of the suggestion that things might be some way, although we cannot so much as consider what that way might be? This sounds like nonsense; and a nonsensical possibility is no possibility at all.

Readers familiar with Berkeley will recognize this line of reasoning. Berkeley dismisses the possibility of a material world, a world of objects existing mind-independently, on the grounds that we cannot so much as entertain thoughts concerning such a world. If we cannot entertain thoughts concerning *X*, then plainly we could have no reason to think *X* exists or might exist: talk of *X* is empty.

The situation we have been envisaging, however, is not one in which we endeavor to think the unthinkable, but one in which we acknowledge our fallibility, recognizing that we could be wrong about almost anything without being in a position to entertain thoughts as to how things actually are. Perhaps this is all a realist needs: it is possible that reality (or a significant portion of reality) is not just unknown, but unknowable by us owing to our circumstances. Or perhaps a realism of this sort leaves behind the traditional impetus for realism. In either case, Putnam's reflections push realists – and

their bedfellows, the skeptics – to examine their fundamental assumptions. For that, realists and anti-realists alike should be grateful.

Ontological pluralism

We have been focusing on Putnam's contention that the world is not mind-independent: "the mind and the world jointly make up the mind and the world." But there is another dimension to Putnam's dissection of metaphysical realism (see Putnam 1987, lecture 1). A metaphysical realist regards the world as possessing a definite character quite independently of our ways of thinking about it. We represent this character in various ways: truly or falsely, subtly or clumsily. Our everyday beliefs about the world represent it as being one way, for instance; the sciences represent it differently. The "scientific image" and the everyday, "manifest image" of the world appear in various ways to be at odds. (Talk of scientific and manifest images originated with Wilfrid Sellars (see SELLARS); see Sellars 1963, ch. 1.) The surface of the desk at which I am sitting appears smooth and continuous. Physics, however, tells us that the desk is a cloud of particles, widely spaced and in constant motion. Which description of the desk is the correct one? Perhaps the apparent desk, the desk of the manifest image, is a *mere* appearance.

This is the reaction of the metaphysical realist. The realist starts with the idea that the world is a single definite way. We can describe the world in many different ways; some kinds of description capture the world better than others, however. My ordinary description of my desk, for instance, is at best a crude approximation. Taken literally it is false. We can edge closer to the truth by turning to physics. When we do, we learn that the world contains no desks, only clouds of invisible particles.

Can we make sense of this picture? Return to my desk. How many things are stacked on it? An answer to this question will depend on how we decide to count. We could, for instance, count pencils, pens, books, and memos. We could just as easily count pages of books, parts of pens and pencils. Or we could count particles of which all these things are composed. What is *the* correct way to count? What is *the* correct answer to the question, how many things are on my desk? Such questions are wrong-headed. There are *many* ways to sort objects on my desk, *many* correct answers to the original question. The contents of my desk can be "carved up" in different ways. How we do so depends on *us*: our aims or purposes.

We deploy systems of concepts in representing the world. The metaphysical realist sees these concepts as matching well or badly what is "out there." But this is the wrong model. The concepts we use determine, rather than merely reflect, what is out there, at least in the sense that they determine objects' boundaries, hence what is to count as an object. It is not that there are no divisions in nature, but that there are too many. Our concepts, or rather systems of concepts, make some of these divisions salient. In saying how the world is, we invoke one or another conceptual system or scheme. Which scheme we invoke depends in part on features of us, our needs and interests. As we learn more and as our needs and interests change, our conceptual schemes evolve.

Suppose this is right. It is then hard to see how we could make sense of talk of a world independent of any conceptual scheme (Kant's noumenal world, the "thing in itself"). We could have no way of describing that world, no way of thinking it: describ-

ing and thinking involve representing in terms of a conceptual scheme. Consider a map of the surface of the Earth. We can depict the Earth's surface by means of a Mercator projection, a Peterson projection, a spherical projection. Imagine someone dissatisfied with these insisting that the Earth be depicted using no projection at all! This is what the metaphysical realist demands for representations of reality in general: a representational system or conceptual scheme that is utterly transparent. But a transparent scheme is no scheme at all.

Once you accept that there can be no sense in talk of a scheme-independent world, you may be moved to ask how competing schemes could be evaluated. Again, this is the wrong question. Just as Mercator projections and Peterson projections do not compete, so our modern scientific scheme does not compete with our everyday conception of the world. Both are entirely satisfactory on their own terms, both provide perfectly adequate depictions of our world. An everyday description can be wrong; I may falsely believe that there is a desk in my office. But if this belief is false it is not because science tells us that there are no desks (only clouds of particles). Ontology – what there is – is relative to a conceptual scheme. Desks do not figure in the ontology of the physicist, but this does not mean that desks are mere appearances. Insofar as we find it useful, or unavoidable, to deploy a conceptual scheme in which desks have a role, the ontological legitimacy of desks is assured. The metaphysical realist hankers after a single ontology: *the* ontology of the world. Instead, Putnam insists, we should embrace ontological pluralism: what there is depends in part on schemes or systems of concepts we find it convenient to deploy.

Externalism again

In assessing Putnam's attack on metaphysical realism, we have been granting externalism, the view that what we mean and what our thoughts concern depends in part on our circumstances. Putnam's defense of externalism relies heavily on Twin Earth cases: we imagine agents who are intrinsically indiscernible and yet whose utterances, and thoughts those utterances express, differ in significance.

Part of the idea here is that the projective character of thoughts – what is often called their *intentionality* – is due, not to intrinsic features of those thoughts, but to matters external to thinkers. Another of Putnam's examples nicely illustrates the point. Suppose you form a mental image of a particular tree, one in a nearby park, for instance. Now imagine an Alpha Centaurian, Fred, who lives on a planet barren of vegetation and so knows nothing of trees. One day Fred spills some paints that purely by chance form a design that you would regard as perfectly realistic representation of the tree in the park. Later, reflecting on the spilled paint, Fred forms a mental image indistinguishable intrinsically from your tree image. Is Fred imagining the tree in the park? That seems unlikely. If Fred is imagining anything, he is imagining a design produced by spilled paint. What accounts for the difference? The images are intrinsically alike, so the difference must lie elsewhere. Perhaps the difference stems from your being causally related to the tree in the park, in a way Fred is not. Your image of a tree *projects* to that tree because you stand in an appropriate causal relation to the tree; Fred's image projects, not to the tree but to the spilled paint, because the spilled paint, not the tree, plays the required causal role.

A view of this sort reverses the metaphor of projection. Thought does not project from the "inside out," but from "outside in." Must we go along? Perhaps not. Perhaps we can account for the projective character of thought by reference to intrinsic features of agents; perhaps projection *is* "inside out." How, then, could we accommodate Twin Earth cases? Suspend doubt for a moment and pretend that the projectivity of a thought is like the beam of a flashlight radiating outward. What the beam of the flashlight illuminates depends both on the nature of the beam, as determined by intrinsic features of the flashlight, and on what happens to be "out there" to be illuminated. Just as flashlights on Earth illuminate water (H_2O), and flashlights on Twin Earth illuminate twin water (XYZ), so thoughts on Earth project to water, and thoughts on Twin Earth project to twin water. The moral? Twin Earth cases do not establish that the projective character of thought is due to incoming causal chains, then, only that what thoughts designate depends in part on the circumstances of thinkers.

Granted it is silly to compare the projectivity of a thought to the beam of a flashlight. Nevertheless it may be possible to base an account of the projective aspect of thought on intrinsic features of agents. Agents and their states of mind possess dispositionality, and dispositions are inherently projective. Locke's example of a lock and key illustrate the idea. The key is *for* locks of a certain sort, and not for others. This is so even if no such lock has been manufactured (or if the lock the key fits is destroyed). The key is disposed to open one lock, but not another: the key *projects* to one lock, but not to another. The key's so projecting does not depend on the key's having been in causal contact with the lock, but solely on intrinsic features of the key (and intrinsic features of the lock).

Imagine now that an agent's states of mind incorporate fine-tuned dispositions (Martin and Heil 1998; Martin and Pfeiffer 1986). Your thoughts of trees, for instance, project to trees, not perhaps because they are caused by trees, but because they dispose you to interact in appropriate ways with trees. To be sure, intrinsically indiscernible thoughts might dispose an inhabitant of Twin Earth to interact with twin trees. This does not show that your thought's projective character comes from the outside, however. The projective character of those thoughts might be "built in" even if objects those thoughts "illuminate" depends on what objects are available to be illuminated.

Consider another much-discussed case (Davidson 1987). Don is wading through a swamp in a thunderstorm. Suddenly, a bolt of lightning reduces Don to a pile of ashes and simultaneously reconstitutes a nearby tree stump into a "molecular duplicate" of Don. Suppose that the molecular duplicate, Swampman, functions just as Don did prior to his sudden demise, and in particular Swampman has thoughts, images, and memories intrinsically indiscernible from Don's. Swampman has many false memories. He seems to remember his twelfth birthday party, but he is in fact only a few minutes old. What of Swampman's other thoughts and images, however? Swampman entertains thoughts and forms images intrinsically indiscernible from Don's thoughts and images of trees, stars, water, and so on. Ought we to say that these thoughts and images are empty of significance until Swampman comes into causal contact with trees, stars, and water? Only someone with a prior commitment to an "outside-in" conception of thought would say so. Swampman's mental condition includes finely-tuned dispositions that undergird the projective aspect of his thoughts. Of course, what those thoughts project *to* depends in some measure on what is "out there." This is the lesson

of Twin Earth. But this need not lead us to imagine that the projectivity of thought must be explained by incoming causal chains.

Putnam's significance

This entry provides only the briefest introduction to one corner of Putnam's philosophical work. I have expressed reservations concerning two themes that have proved especially influential in recent philosophy. These critical comments afford, at best, only hints as to where a reader might disagree with those doctrines. Putnam's work is wide-ranging, rich, and interconnected in a way that undercuts piecemeal criticism. In attacking one Putnam thesis, a critic risks assuming positions that Putnam elsewhere rejects. Perhaps I have said enough to make it clear that Putnam's work is deeply insightful, penetrating, and synoptic. Even when Putnam self-confessedly goes up a blind alley, it is worth following him for the sake of observing some topic in a new and revealing light. (Besides, Putnam's blind alleys are more interesting than the well-trodden paths of other philosophers.) Hilary Putnam is one of a handful of philosophers who have individually shaped the fundamental character of contemporary philosophy.

Bibliography

Works by Putnam

1971: *Philosophy of Logic*, New York: Harper and Row.
1975a: *Mathematics, Matter, and Method,* Philosophical Papers vol. 1, Cambridge: Cambridge University Press.
1975b: *Mind, Language, and Reality*, Philosophical Papers vol. 2, Cambridge: Cambridge University Press.
1978: *Meaning and the Moral Sciences*, London: Routledge and Kegan Paul.
1981: *Reason, Truth, and History*, Cambridge: Cambridge University Press.
1983: *Realism and Reason*, Philosophical Papers vol. 3, Cambridge: Cambridge University Press.
1987: *The Many Faces of Realism*, 1985 Paul Carus Lectures, La Salle, IL: Open Court.
1988: *Representation and Reality*, Cambridge, MA: MIT Press.
1990: *Realism with a Human Face*, ed. J. Conant, Cambridge, MA: Harvard University Press.
1992: *Renewing Philosophy*, Cambridge, MA: Harvard University Press.
1994: "Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind," *Journal of Philosophy* 91, pp. 445–517.
1995: *Pragmatism: An Open Question*, Oxford: Blackwell Publishers.

Works by other authors

Boolos, G. (ed.) (1990) *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge: Cambridge University Press.
Clark, P. and Hale, B. (eds.) (1994) *Reading Putnam*, Oxford: Blackwell Publishers.
Davidson, D. (1987) "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association* 60, pp. 441–58.
Dennett, D. (ed.) (1987) *The Philosophical Lexicon*, Newark, DE: American Philosophical Association.

Emerson, R. W. (1940) "Self-reliance," in *The Complete Essays and Other Writings of Ralph Waldo Emerson*, ed. B. Atkinson, New York: The Modern Library.
Heil, J. (1998) "Skepticism and Realism," *American Philosophical Quarterly* 35, pp. 57–72.
Lewis, D. (1972) "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy* 50, pp. 249–58.
Martin, C. B. and Heil, J. (1998) "Rules and Powers," *Philosophical Perspectives* 12, pp. 283–312.
Martin, C. B. and Pfeiffer, K. (1986) "Intentionality and the Non-psychological," *Philosophy and Phenomenological Research* 46, pp. 531–54.
Place, U. T. (1956) "Is Consciousness a Brain Process?," *British Journal of Psychology* 47, pp. 44–50.
Ryle, G. (1949) *The Concept of Mind*, London: Hutchinson.
Sellars, W. (1963) *Science, Perception, and Reality,* London: Routledge and Kegan Paul.
Smart, J. J. C. (1959) "Sensations and Brain Processes," *Philosophical Review* 68, pp. 141–56.
Wittgenstein, L. (1953) *Philosophical Investigations,* trans. G. E. M. Anscombe, Oxford: Blackwell Publishers.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

33

David M. Armstrong (1926–)

FRANK JACKSON

David Armstrong's many major contributions are focused in traditional epistemology and metaphysics. He offers comprehensive accounts of what there is, its nature, and how we know about it. He is a "system builder." His work is informed by the conviction that philosophers must take very seriously the teachings of science. He is a realist: about mental states, about properties, about laws, and about singular causation. Indeed, on almost any philosophical topic, if there is a realist position available, Armstrong will occupy it. Also, he seeks what is now often called "the view from nowhere." He is opposed to the idea that there may be different, equally legitimate but, to one degree or another, incommensurate, views of how things are from one or another perspective. Or, as it is sometimes put, he denies that there are different kinds of being or of truth.

Materialism about the mind

Armstrong is probably best known for *A Materialist Theory of the Mind*. His theory is commonly known as central state materialism or as the causal theory of mind.

Armstrong started his philosophical life as a behaviorist but, partly as a result of the influence of J. J. C. (Jack) Smart, moved to the view that mental states are states of the central nervous system, and more especially the brain. Armstrong develops his central state version of the identity theory by first arguing that the concept of a mental state *M* is the concept of a state that plays a distinctive causal role that connects stimulus, behavioral response, and other mental states. Thus, to give the rough idea, pain is the state typically caused by bodily damage, and typically causing a desire that it itself cease and a behavioral response that tends to, or is believed to, minimize the damage. Obviously, an account of this kind is exactly what evolutionary considerations would suggest. In similar fashion, belief is a state induced by subjects' environments that tends to make them behave in ways that realize what they desire if what they believe is true. Armstrong sees two major advantages of this kind of view over behaviorism. First, it allows mental states to be causes of behavior. Secondly, by bringing in reference to other mental states, it allows for suitably complex accounts of the connections between mental states and behavior. It is notorious that there is no simple one-to-one matching of mental states and behavior. What you do when you think it is about to rain depends,

inter alia, on whether you want to stay dry, on where you think the umbrella is, on whether you think you are Gene Kelly, and on how cold you feel.

On the central state theory, to ask after the identity of a given mental state *M* is to ask what state plays the distinctive, causally intermediate role assigned by the concept of *M*. Armstrong argues that, for each mental state, it will turn out to be some state or other of the brain that plays the role in question. He concludes, therefore, that, as an empirical matter of fact, mental states are identical with brain states.

These identities will be contingent because which brain states play which roles is a contingent matter. Some have objected that there are no contingent identities: everything is necessarily identical to itself, and that what Armstrong (and Smart) should say is that the identities are a posteriori. In fact, they hold that the identities are both contingent and a posteriori, but the objection to the contingent identity part of their theory is a misunderstanding. All they mean is that sentences of the form "*M* is *B*" are contingent, in the same way that "Red is the color of bullfighters' capes" clearly is.

A second misunderstanding is over Armstrong's stance on the possibility that quite different states might play the causal role distinctive of pain in different species and, maybe, in different members of the one species. It is often objected that identity theorists are committed (implausibly) to pain being the same state in everything that experiences pain. But consider the following parallel. The most dangerous virus for dogs is different from the most dangerous virus for people, and the reason for this is that the kind that plays the relevant role in dogs differs from the kind that plays the relevant role in people. Nevertheless, we can, and do, *identify* the most dangerous virus for dogs and for people – or anyway the experts do it for us.

A more pressing question is whether Armstrong (and Smart) should have said that mental states are constituted by, rather than identical with, brain states. The relation between a table and the parts that make it up is one of constitution, not identity. Because the life histories of the table and its parts differ – for example, the parts typically come into existence somewhat earlier than the table – Leibniz's Law means that the relation cannot be one of identity; it must be constitution. (A separate question is whether this relation of constitution can be analyzed in terms of identity between temporal parts of the table and temporal parts of various aggregations of parts.) Similarly, it may well be that Armstrong (and Smart) should, strictly, say that mental states are constituted by brain states, not that they are identical to them.

Armstrong's central state view of mind is sometimes contrasted with the kind of functionalist theory of mind associated with the early Hilary Putnam (see PUTNAM). They both agree in giving functional roles a central role in the theory of mind. This is because Armstrong's causal roles can equally be described as functional roles. The stimuli that Armstrong talks of are inputs, as functionalists say it, and the behavioral responses are outputs, as functionalists say it. There are two big differences, though. Armstrong thinks of the mental states as the occupants of the functional roles, as the states that are suitably interconnected to inputs, outputs, and other, internal mental states. Putnam thinks of them (or thought of them when he was a functionalist) as the functional roles themselves. And, secondly, the functional roles in Armstrong's theory are those sometimes called "common sense." Their inputs and outputs are described in terms familiar to us all: rain, umbrellas, movements that lead to beer inside the mouth,

etc. In Putnam's version of functionalism, though not in all versions of functionalism, the inputs and outputs are thought of as internal ones.

Perception, sensations, belief, knowledge

Armstrong's *Perception and the Physical World* is an argument for direct realism in perception. He argues that we are directly acquainted with independently existing physical objects in perception. The distinctive feature of his argument is the way it is founded on an analysis of perception and perceptual experience in terms of the acquisition of belief through the operation of one's sense organs. This makes good sense of the central biological function of perception, which is the acquisition of belief about what is going on around and inside one. An obvious question for Armstrong's account is raised by the fact that the very same belief, say, that it is raining outside, can be acquired in very different ways through very different perceptual experiences. You might, for instance, see that it is raining, be told that it is raining, read on a computer screen that it is raining, or hear that it is raining. Perhaps the most plausible way of approaching this problem is in terms of the distinctively different clusters of beliefs in each case. In none of these cases, does the belief that it is raining come "by itself"; rather, it comes as an integral part of a whole cluster of beliefs, and the clusters are different in, and distinctive of, each case.

In *Bodily Sensations*, Armstrong gives an account of somatic sensations in terms of perception of one's own body. A sensation is an experience of perceiving that one's body is in such and such a state, an experience which may or may not be veridical. For example, a feeling of warmth is the putative perception that a part of one's body is warm. In the case of certain sensations, the putative perception is accompanied by a characteristic attitude. Pain, for example, is the putative perception that there is something amiss with part of one's body, accompanied by an immediate dislike of this putative perception.

Armstrong's treatment of belief follows a suggestion of F. P. Ramsey's that belief is like a map by which we steer. Inside our heads is a master map that moves us through the world in such a way that what we desire is achieved to the extent that the map is correct, and individual beliefs are thought of as sub-maps of the master map. This approach to belief is now a standard alternative to the internal sentence theory of belief supported by language of thought theorists.

His account of knowledge is a reliabilist one. Knowledge necessarily involves true belief: if *S* knows that *P*, then *S* truly believes that *P*. But not all true belief is knowledge; the truth of a belief may be an accident, and how can getting it right by accident be *knowledge*? Armstrong's suggestion, roughly, is that *S*'s true belief that *P* is knowledge if it is a reliable sign that *P*. Here he differs from the tradition that requires that one's belief be justified in order to count as knowledge.

Time and action

Armstrong holds a temporal part, or stage, metaphysics. Identity over time is a matter of having parts or stages at the times in question. I was at the Melbourne Test when "Typhoon" Tyson took 7 for 27 because a certain person-stage with the right

connections to the person-stage writing these words was present at that test match. Armstrong's main contribution to the debate is one of the very first discussions of the famous rotating homogeneous cylinder/disk/sphere example. He argues that the example shows the *conceivability* of a conception of identity through time not framed in terms of temporal stages, but that, nevertheless, the temporal stage account of identity through time is in fact correct. What makes it true, on his view, that such an object is rotating are the dependencies between different stages.

Armstrong was also one of the first, with Brian O'Shaughnessy, to argue that if one acts, one must have tried to act, and that this is the essence of truth in the old volitional theory of action.

Universals, laws, causation, possibility, and states of affairs

Truth-makers play a crucial role in Armstrong's later philosophizing. The basic idea is that if some sentence or proposition is true, there must be something that makes it true; similarly, if some predicate applies to something, there must be something that makes it true that the predicate applies. You cannot say that the word "square" applies to *A*, and that that is *all* there is to say. There must be something about *A* that makes it true that the word applies to it, that *A* satisfies it. In Armstrong's hands, the truth-maker principle, as he calls it, is more than the widely accepted supervenience of truth and satisfaction on nature. Supervenience says that if a sentence is true in one situation and false in another, and if a predicate is satisfied by one thing but not by another, the situations and things must differ in nature. The truth-maker principle goes further. It says that there must be something that makes – necessarily makes – the true sentence true and the satisfied predicate satisfied.

Armstrong holds that what makes it true that predicates apply to particulars are the properties or universals that the particulars possess. In keeping with his realist leanings, these universals exist independently of the classifications that we find natural. They are in nature. Secondly, they are not to be reduced to sets, or to resemblances between particulars. Armstrong is not a nominalist. He argues, in particular, that nominalists cannot handle the famous "one over many" problem, the problem of what unifies things that share a property. Thirdly, there are no uninstantiated universals; every universal is possessed by at least one thing. In this sense, he is with Aristotle and not Plato. He regards the Platonic view that there are uninstantiated properties or universals as an unmotivated ontological extravagance. Fourthly, there is not a one–one relation between properties and predicates: one and the same universal or property may be the truth-maker for a number of different predicates. To illustrate: suppose that *U* is a universal and that "*A*" is a predicate that says that something is *U*, and "*B*" is some quite different predicate. Surely, "*A* or *B*" might be true of something which is *U* simply because it is *U*. We are not required to postulate an extra property just because "*A* or *B*" is a distinct predicate from "*A*." Also, there may be properties for which there is no predicate. Finally, which universals or properties there are is an a posteriori matter to be settled by total science. Philosophy tells us that there must be truth-makers for true predications, but what they are is ultimately a matter for science broadly conceived.

Armstrong argues strongly against Humean and neo-Humean accounts of laws. For him, no facts about regularities, however tricked up, can ever add up to lawfulness

proper. What then must be added to a regularity to get a law? His answer is that what distinguishes the universal statements of the form "Every *F* is *G*" that express laws of nature – that are nomic or nomological – from those that express accidental regularities is that, roughly, the laws correspond to relations of nomic necessitation between universals. In its simplest version, the idea is that "Every *F* is a *G*" is a law if and only if Fness necessitates Gness. But more detailed accounts would need to advert to his metaphysics of states of affairs, mentioned briefly below, and to his treatment of laws that do not fall obviously into the "Every *F* is *G*" mold, derived laws, and laws that have no instances (for example, concerning motion in the absence of gravity).

This account of laws is, obviously, strongly anti-Humean. Armstrong's account of causation is equally counter to the tradition that comes to us from Hume, and in three respects. First, Armstrong insists that causation is singular in that it is a non-relational property of a sequence (see ANSCOMBE). Secondly, he holds that the connection between causation and law is a posteriori. He denies, that is, that it is a priori that any singular causal sequence falls under some law. He does, though, allow that it may well be that some or all causal sequences are identical, as an a posteriori matter, with the instantiation of a law. Finally, he holds that we sometimes directly perceive singular causal connections. Here he is going against a widely held view, even among those who would not describe themselves as Humeans. Many who agree with him that causation is more than sequence suitably constrained think, nevertheless, that sequence is all we literally perceive. We do not see that *X* caused *Y*; we infer it. Sometimes their argument for this view is that a non-causal sequence can seem as causal as can be, as Piaget's famous experiments tell us. Armstrong rightly points out that this only shows that illusion is possible, and the possibility of illusion concerning a feature does not show that we do not literally perceive it when all goes well. However, there is a stronger argument. It is hard to identify the causal role that singular causation plays in its alleged perception. When I see that something is square, its squareness plays a role in inducing my perceptual experience. This seems crucial to its being correct to say that I perceive its squareness. But what role does singular causation play that might mirror the role squareness plays? All the causal work seems to be being done by the sequence *per se*.

Armstrong's account of possibility is a combinatorial one, drawing on his realism about universals. We can think of how things are as a vast, complex arrangement of particulars and universals. The various possibilities can then be thought of as all the combinations and recombinations of these particulars and universals according to various rules for combining particulars and universals. Thus, to give the barest bones of the idea, suppose that there is in fact charge *X* at point *y*, and charge *U* at point *v*. What makes it possible that there be charge *X* at *v*, and charge *U* at *y*? His answer is the fact that putting *X* with *v*, and putting *U* with *y*, does not violate the rules of combination.

In his most recent book, *A World of States of Affairs*, Armstrong argues that the best way to bring his ideas on universals, laws, truth-making, and possibility together is by adopting a metaphysics of states of affairs. For example, universals – the key to his account of laws – turn out to be types of states of affairs. In any case, for Armstrong, the world is not the aggregation of all the things there are. It is the aggregation of all the states of affairs there are, where states of affairs are things-having-properties. His

view is essentially the same as Wittgenstein's in the *Tractatus*, namely, that the world is the totality of facts, not of things.

Bibliography of works by Armstrong

1960: *Berkeley's Theory of Vision*, Melbourne: Melbourne University Press.
1961: *Perception and the Physical World*, London: Routledge and Kegan Paul.
1962: *Bodily Sensations*, London: Routledge and Kegan Paul.
1968: *A Materialist Theory of the Mind*, London: Routledge and Kegan Paul.
1973: *Belief, Truth and Knowledge*, Cambridge: Cambridge University Press.
1978: *Universals and Scientific Realism*, 2 vols., Cambridge: Cambridge University Press.
1983: *What is a Law of Nature?*, Cambridge: Cambridge University Press.
1984 (with Norman Malcolm): *Consciousness and Causality: A Debate on the Nature of Mind*, Oxford: Blackwell Publishers.
1989a: *A Combinatorial Theory of Possibility*, Cambridge: Cambridge University Press.
1989b: *Universals: An Opinionated Introduction*, Boulder, CO: Westview Press.
1997: *A World of States of Affairs*, Cambridge: Cambridge University Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

34

Noam Chomsky (1928–)

PETER LUDLOW

Noam Avram Chomsky, born in Philadelphia, Pennsylvania, received his Ph.D. in linguistics from the University of Pennsylvania in 1955. Since 1955 he has taught at MIT, where he currently holds the position of Institute Professor. Chomsky gained the attention of philosophers early on in his career by the introduction of mathematical/logical tools for the description of linguistic phenomena. In this respect his early work was influenced by figures such as Nelson Goodman and W. V. Quine, both of whom are thanked in the introduction to his *Syntactic Structures* (1957). Nevertheless, Chomsky's principal philosophical significance relates to his rejection of the approach to language and mind taken by Quine and many other analytic philosophers. Indeed, Chomsky has been a direct participant in several key philosophical debates in the last half century, taking issue with interlocutors such as Quine, Donald Davidson, Hilary Putnam, Saul Kripke, and John Searle on the nature of language and mind.

In the view of many analytic philosophers, language is a social object that has been established by convention for purposes of communication. Chomsky's take is different: the conception of language as an external social object is unfruitful (if not incoherent), and the only plausible strategy for the empirical scientist is to view language, or rather, the language faculty, as a natural object that is part of our biological endowment.[1] The exact nature of this picture has evolved since the 1960s, and it has taken the form of a "principles and parameters model," which can be viewed in the following way: think of the language faculty as being a largely pre-wired mechanism with a set of switches (parametric settings) which can be set in various ways depending upon the environmental setting into which the language-learner is born. The task of the linguist is to study this mechanism, to deduce its initial state, and to understand what the possible parametric settings are, that is, to determine precisely what variation is allowed by the language faculty.

Chomsky (1986a) introduces the terms "I-language" and "E-language" to distinguish his general thesis about the language faculty from the loose collection of theories about language that hold that it is social or external. The I-language/E-language distinction is useful, since it highlights the idea that the object of study in linguistics is "internal" in a sense, and is not directly concerned with "external" phenomena like written corpora of data. Chomsky is thus opposed to the conception reflected in the definition given by the American linguist Leonard Bloomfield: language is "the

totality of utterances that can be made in a speech community." For Chomsky, I-languages are "in the mind, ultimately the brain."

To highlight the difference between these two approaches, consider the two different pictures of linguistic rules that emerge. Traditional grammarians (citing conventions and common practice for written English) give us superficial rules such as "Do not end a sentence with a preposition" or "Use 'whom', not 'who' when the pronoun has accusative or dative case." On the other hand, generative grammarians like Chomsky note that there are more subtle and interesting linguistic rules which go unnoticed by the traditional grammarian but which seem to be employed by a broad class of speakers. For example, no native speaker of English would recognize (1) (below) as a well-formed question in English, even though there are seemingly similar structures like (2) that *are* quite acceptable to language users:

1 *Who did John see the boy that Bill hit?
2 Who did John say that Bill hit?

The account for the difference in these cases is subtle, and the details of the explanation have changed as generative grammar has evolved.[2] The fact remains, however, that native speakers of English know that (2) is acceptable and that (1) is not, and further it is clear that no one is taught to have this preference. Whatever rules account for the judgments about (1) and (2) they are far more subtle than the usual prescriptive rules.

Similar considerations apply to the following examples, discussed in Chomsky (1986b; see also 1982, 1986a).

3 John filed every letter without reading it.
4 What letter did John file without reading it?

Somehow speakers of English know that if we delete the pronoun "it" in these two sentences (as in (3′) and (4′)) the effects on meaning are different.

3′ John filed every letter without reading.
4′ What letter did John file without reading?

(4′) is ambiguous in a way that (3′) is not. Both (3′) and (4′) have the meaning in which the filing was done without some (unspecified) reading taking place, but (4′) also preserves the most salient possible meaning of (4): it can still be understood as asking what letter John filed without reading *it* – the filed letter. Clearly, no one taught us this, there is no convention (tacit or otherwise) to use language in this way, and no prescriptive grammarian ever stipulated that we should interpret these sentences in this manner. But, just as clearly, these facts describe the linguistic competence of a large class of individuals.

This is just one of the problems for traditional grammars and for the more general assumption that languages are objects that are established by convention. Our best attempts to stipulate the rules – or to make explicit the conventions – just scratch the surface about our linguistic competence. In Chomsky's words,

> Traditional grammars do not describe the facts of language; rather, they provide hints to the reader who already has, somehow, the requisite "notion of structure" and general conceptual resources, and can use the hints to determine the expressions of the language and

> what they mean. The same is true of dictionaries. . . . Traditional grammars and dictionaries, in short, presuppose "the intelligence of the reader"; they tacitly assume that the basic resources are already in place. (1994b: 160)

Likewise, institutions such as the Académie Française do not stipulate as much as they think they do. At best they give some superficial rules of thumb for proper linguistic behavior regarding French. They cannot even begin to cover the range of facts of interest to practicing linguists.

Indeed, for a generative linguist, the traditional notion of a language like French or German is suspect at best (see Chomsky 1980b: ch. 6). In what sense is a "speaker of German" from the Dutch border of Germany and a "speaker of German" from Bavaria speaking the same language? (Especially given that their languages are not mutually intelligible?) The fact that we say these individuals speak the same language is more of a political decision than anything else, and indeed an individual raised in northern Germany and an individual raised in The Netherlands may find that their languages are more mutually intelligible than the two aforementioned German citizens do. Saying that the two German citizens speak the same language is at best a loose way of talking about some contextually relevant (and certainly political rather than linguistic) similarities. Chomsky (1994b) compares it to saying that two cities are "near" each other; whether two cities are near depends on our interests and our mode of transportation and virtually not at all on brute facts of geography. The notion of "same language" is no more respectable a notion in the study of language than "nearness" is in geography. Informally we might group together ways of speaking that seem to be similar (relative to our interests), but such groupings have no real scientific merit. As a subject of natural inquiry, the key object of study has to be the language faculty and its set of possible parametric variations.

One might think it possible to retreat slightly by giving up on the idea of an E-language and endorsing a notion of E-dialect or E-idiolect, but even this retreat will not save the language-as-external-object position, according to Chomsky. Considerations that make it arbitrary when to say that two individuals speak the "same language" also apply to saying when they speak the "same dialect." Furthermore, we have no way of identifying the linguistic forms that would be part of a given individual A's E-idiolect. In the first place, A speaks in different ways with different groups of individuals (say A uses a different vocabulary among philosophers than among family members) and indeed at different stages of life (contrast A's use of language at age 3 and age 30). Do all of these ways of speaking count as being part of the same idiolect? What unifies them other than that they are ways in which A happens to have spoken? Still worse, we certainly can't identify A's E-idiolect with some corpora of utterances and inscriptions, for these intuitively include speech and spelling errors. On the basis of what can we say that a given hiccup is an error and not part of the spoken corpus of A's E-idiolect? If we try and identify errors by appealing to A's language community, that lands us back in the problem of individuating E-languages and E-dialects; there is simply no fact of the matter about which language community A belongs to.

On the I-language approach, however, this problem takes the form of a well-defined research project. The idiolect (I-idiolect) is determined by the parametric state of A's

language faculty; the language faculty thus determines A's linguistic *competence*. Speech production that diverges from this competence can be attributed to *performance* errors. Thus, the competence/performance distinction is introduced to illuminate the distinction between sounds that are part of A's grammar and those that are simply mistakes. The E-language perspective has no similar recourse.

For Chomsky, these are among the myriad reasons we have for abandoning the idea that language (as studied by the linguist) is a social object, and adopting the perspective that it is a natural object. But what kind of natural object? Since children acquire their linguistic competence without serious formal training (certainly none that would cover the facts in (1)–(4)) and indeed with impoverished data, Chomsky hypothesizes that there must be an innate language acquisition device which accounts for this competence. The task of the linguist is to learn the initial state of this device, and to determine the possible parametric variations of the device that are brought about by exposure to linguistic data.

This thesis has led to controversy; indeed, it has come to be at the center of recent innateness debates between Chomsky and Piaget, and Quine among others (see QUINE). The debates have turned on whether language acquisition requires a dedicated language faculty or whether "general intelligence" is enough to account for our linguistic competence. Chomsky considers the "general intelligence" thesis hopelessly vague, and argues that generalized inductive learning mechanisms make the wrong predictions about which hypotheses children would select in a number of cases. Consider the following two examples from Chomsky (1975, 1980a).

5 The man is tall.
6 Is the man tall?

Chomsky observes that confronted with evidence of question formation like that in (5) and (6) and given a choice between hypothesis (H1) and (H2), the generalized inductive learning mechanism will select (H1).

(H1) Move the first "is" to the front of the sentence.
(H2) Move the first "is" following the first NP to the front of the sentence.

But children apparently select (H2), since in forming a question from (7) they never make the error of producing (8), but always opt for (9).

7 The man who is here is tall.
8 *Is the man who here is tall?
9 Is the man who is here tall?

Note that this is true despite the fact that the only data they have been confronted with before encountering (7) is simple data like (5) and (6). Chomsky's conclusion is that whatever accounts for children's acquisition of language it cannot be generalized inductive learning mechanisms, but rather must be a system with structure-dependent principles/rules. In effect, one has to think of the language faculty as being a domain-specific acquisition *module*.[3]

Obviously the distinction between I-language and E-language puts Chomsky at odds with a number of philosophers on the nature of language, but it also leads to a number of subsidiary philosophical disputes, not least of which are those disputes that are

driven by questions about the nature of rules and representations (or principles and parameters) in cognitive science.

For example Quine extends his "gavagai" argument and attendant skepticism about meanings to similar skepticism about grammatical rules. At the core of Quine's worry is the idea that if several rule systems are consistent with the linguistic behavior of an individual, then there can be no fact of the matter about what set of rules is actually being employed (see QUINE). Chomsky (1969, 1975, 1980) has made several responses to this argument. In the first place, Chomsky takes Quine's argument to be a rehash of the standard scientific problem of the underdetermination of theory by evidence. So, for example, even if there are several grammars that are consistent with the available linguistic facts (not linguistic behavior, for Chomsky, but intuitions about acceptability and possible interpretation) we still have the additional constraint of which theory best accounts for the problem of language acquisition, acquired linguistic deficits (e.g. from brain damage), linguistic processing, etc. In other words, since grammatical theory is embedded within cognitive psychology, the choice between candidate theories can, in principle, be radically constrained. But further, even if we had two descriptively adequate grammars, each of which could be naturally embedded within cognitive psychology, there remain standard best theory criteria (simplicity, etc.) which can help us to adjudicate between the theories.

A more recent assault on rules and representations has come from Kripke's reconstruction of Wittgenstein's private language argument (see KRIPKE). According to that argument, there can be no fact of the matter about what rules and representations a system of unknown origin may be following. Kripke concludes:

> *if* statements attributing rule-following are neither to be regarded as stating facts, nor to be thought of as *explaining* our behavior . . . it would seem that the use of the idea of rules and of competence in linguistics needs serious reconsideration, even if these notions are not rendered meaningless. (1982: 31 n. 22)

Chomsky's initial (1986b: ch. 4) response to the Kripke/Wittgenstein argument appears to be that there is a fact of the matter about what rules a computational system is operating on, but in more recent articles (1993, 1994b, 1995a) he has argued that the Kripke/Wittgenstein argument applies only to artifacts and not to natural objects. That is, computers are artifacts – the products of human intentions – and hence there is no fact about their design that exists apart from those intentions. The principles and parameters of the language faculty, on the other hand, are embedded within cognitive psychology and ultimately facts about human biology. Therefore the structure of the language faculty is no less grounded than, for example, the human genome.

Chomsky has also clashed with Searle over the possibility of rules in cognitive science that are "in principle" inaccessible to consciousness. Can there be aspects of the mental which are not "in principle" accessible to consciousness? Searle argues that there cannot be (see SEARLE). Chomsky (1990, 1994a) argues that the notion of "in principle" in Searle's argument is vacuous. For example, what evidence is there that the grammatical principles governing our judgments about examples (1)–(9) can't be accessible to consciousness? Is it a law of logic that there could not be a species with a language faculty just like ours but with full conscious access to its principles and

parameters? Chomsky also notes that Searle must introduce the notion of "blockage" to cover those cases in which an individual, perhaps through brain damage, is able to correctly solve a problem, but be unable to say how it was solved. On Searle's theory, such a person has "in principle" access but suffers from "blockage." But Chomsky observes that it is entirely arbitrary as to what counts as blockage and what counts as in principle inaccessibility (e.g. perhaps an unfortunate mutation blocked our access to the language faculty). Accordingly, Chomsky argues that such notions have no role in naturalistic inquiry into the nature of the mental (and indeed, cognitive science rightly ignores such notions).

For Chomsky, it is not enough to defend the idea that rules and representations (principles and parameters) be a part of our naturalistic investigation into the mind, their character must also be *individualistically* determined. That is, there is a brute fact about the state of an individual's language faculty and that fact is determined in turn by facts about the individual in isolation, not by the environment in which the individual is embedded. The thought is that if the language faculty is part of our biological endowment, then the nature of the representations utilized by the language faculty are fixed by our biology and are not sensitive to environmental issues such as whether we are moving about on Earth or Twin Earth.

This appears to put Chomsky on a collision course with figures such as Tyler Burge, who argues in "Individualism and Psychology" that the content of the representations posited in psychology are determined at least in part by environmental factors. If the notion of content involves externalist or environmental notions, then Chomsky is dubious that it can play an interesting role in naturalistic inquiry in cognitive psychology. Furthermore, since for Chomsky "the mental" is simply an aspect of the natural world that is investigated by sciences such as cognitive psychology (see Chomsky 1994a), the nature of the mental itself must be individualistically and not environmentally determined. Thus Chomsky (1993, 1995a) rejects the contentions of figures such as Putnam ("The Meaning of 'Meaning' "), Burge ("Individualism and the Mental"), and Davidson ("Knowing One's Own Mind") that the contents of our mental states are environmentally determined. More accurately, he dismisses the talk of "contents" as ill-defined. Indeed, he is dismissive of the thought experiments (Putnam's "water/twater," Burge's "tharthritis," and Davidson's "Swampman") that purport to support externalism, and suggests that they reflect philosophical prejudice more than any genuine facts about the mind/brain.

If environmentalism is to be rejected in psychology, then it naturally must be rejected in semantics as well. That is, if the task of the linguist is to investigate the nature of I-language, and if the nature of I-language is a chapter of cognitive psychology, and if cognitive psychology is an individualistic rather than a relational science, semantics will want to eschew relational properties like reference (where reference is construed as a relation between a linguistic form and some object in the external environment). Thus Chomsky (1981, 1995b) rejects the notion of reference that has been central to the philosophy of language since about 1970, characterizing it as an ill-defined technical term (certainly one with no empirical applications), and suggesting that in the informal usage of "refer," individuals refer but linguistic objects do not.

It also follows that semantic theories that employ the technical notion of reference should be rejected in favor of semantical theories which do not purport to state lan-

guage/world relations or, following Chomsky (1975a), in favor of a Wittgensteinian approach in which there is no semantics *per se*, but rather one in which the lingustic forms are *used* in certain ways.

With this rejection of referential semantics also comes a rejection of any attempt to use the semantics of natural language to gain insights into ontology. It is no good to argue from the structure of language to the existence of events, or plural objects, or times, etc. As Chomsky has argued, there are a number of constructions where the structure of language and the structure of the external world diverge. For example, some noun phrases intuitively have counterparts in the world (for example, the noun phrase "coats in the closet") while others do not ("flaws in the argument"):

> If I say "the flaw in the argument is obvious, but it escaped John's attention," I am not committed to the absurd view that among things in the world are flaws, one of them in the argument in question. Nevertheless, the NP *the flaw in the argument* behaves in all relevant respects in the manner of the truly referential expression the coat in the closet. (1981: 324)

Still more, Chomsky holds that there is a deep reason why our ontology cannot be reflected in natural language: ontology is determined by human intentions, while the representations in the language faculty are naturalisitically determined.

> We do not regard a herd of cattle as a physical object, but rather as a collection, though there would be no logical incoherence in the notion of a scattered object, as Quine, Goodman, and others have made clear. But even spatiotemporal contiguity does not suffice as a general condition. One wing of an airplane is an object, but its left half, though equally continuous, is not. . . . Furthermore, scattered entities can be taken to be single physical objects under some conditions: consider a picket fence with breaks, or a Calder mobile. The latter is a "thing," whereas a collection of leaves on a tree is not. The reason, apparently, is that the mobile is created by an act of human will. If this is correct, then beliefs about human will and action and intention play a crucial role in determining even the most simple and elementary of concepts. (1975b: 204)

The upshot is that pursuing metaphysical questions by appeal to natural language (I-language) is a dead end.

Perhaps less clear are the prospects for recent philosophical attempts to employ the resources of generative grammar in carrying out Davidson's program of defining truth in natural language (see DAVIDSON). Chomsky's view appears to be that everything depends upon how these enterprises are interpreted. If they are taken to be ways of executing a referential semantics, then they are misguided. If the "semantic values" of these theories are taken in a non-referential way then there is presumably room for interesting theorizing.

In this entry I've only scratched the surface of work by Chomsky that is potentially of interest to philosophers (analytic or otherwise). One glaring omission is his writing on social issues (see, for example, Chomsky 1987) and on the media (see Herman and Chomsky 1988). I've also passed over his contributions for formal language theory (Chomsky 1956, 1959) and general issues in epistemology (Chomsky 1981). Finally, there is much that could have been said about his earlier syntactic work (1957, 1965,

1975a) and the influence that it had in the philosophy of language in the 1960s and 1970s. I hope, however, that the forgoing discussion has helped to illuminate some of Chomsky's work and placed it in the context of the debates that have taken place in analytic philosophy since the mid-twentieth century, debates which remain open largely due to his efforts.[4]

Notes

1 Chomsky often suggests that if one digs beneath the surface, one finds that these philosophers (even the behaviorists) are also believers in a language faculty which is part of our biological endowment. See his discussion of Quine in Chomsky 1975b: 198ff. On Chomsky's view, of course, no coherent story can be told without this assumption.

2 In Chomsky 1975b, 1977, for example, the idea is that (1) represents a *subjacency* violation; formation of the question would require the wh-element "who" to move out of both an NP (noun phrase) and an S (clause) without a safe intermediate landing site. For more current accounts of these constructions see Chomsky 1986a, 1995b.

3 However, Chomsky observes that it is not modular in the sense of Fodor's *Modularity of Mind*, but is an acquisition module more in the sense of Gallistel in *The Organization of Learning*.

4 I am indebted to Noam Chomsky, Richard Larson, and A. P. Martinich for comments on an earlier draft of this article.

Bibliography

Works by Chomsky

1956: "Three Models for the Description of Language," *I.R.E. Transactions of Information Theory*, IT-2, pp. 113–24.
1957: *Syntactic Structures*, The Hague: Mouton.
1959: "On Certain Formal Properties of Grammars," *Information and Control* 2, pp. 137–67.
1965: *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press.
1966: *Cartesian Linguistics*, New York: Harper and Row.
1969: "Quine's Empirical Assumptions," in *Words and Objections: Essays on the Work of W. V. Quine*, ed. D. Davidson and J. Hintikka, Dordrecht: D. Reidel.
1971: *Problems of Knowledge and Freedom: The Russell Lectures*, New York: Vintage Books.
1975a: *The Logical Structure of Linguistic Theory*, Chicago: University of Chicago Press. (Originally appeared in unpublished manuscript form in 1955.)
1975b: *Reflections on Language*, New York: Pantheon.
1977: "Conditions on Rules of Grammar," in *Essays on Form and Interpretation*, Amsterdam: Elsevier North Holland, pp. 163–210.
1980a: "On Cognitive Structures and their Development: A Reply to Piaget," in *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, ed. M. Piatelli-Palmerini, Cambridge, MA: Harvard University Press, pp. 35–54.
1980b: *Rules and Representations*, New York: Columbia University Press.
1981: *Lectures on Government and Binding*, Dordrecht: Foris Publications.
1982: *Some Concepts and Consequences of the Theory of Government and Binding*, Cambridge, MA: MIT Press.
1986a: *Barriers*, Cambridge, MA: MIT Press.
1986b: *Knowledge of Language*, New York: Praeger.
1987: *The Chomsky Reader*, New York: Pantheon Books.

1990: "Accessibility 'in Principle'," *Behavioral and Brain Sciences* 13, pp. 600–1.
1992: "Explaining Language Use," *Philosophical Topics* 20 (Spring), pp. 205–31.
1994a: "Naturalism and Dualism in the Study of Language and Mind," *International Journal of Philosophical Studies* 2, pp. 181–209.
1994b: "Noam Chomsky," in *A Companion to the Philosophy of Mind*, ed. S. Guttenplan, Oxford: Blackwell Publishers, pp. 153–67.
1995a: "Language and Nature," *Mind* 104, pp. 1–61.
1995b: *The Minimalist Program*, Cambridge, MA: MIT Press.
1988 (with Herman, E.): *Manufacturing Consent: The Political Economy of the Mass Media*, New York: Pantheon Books.

Works by other authors

George, A. (ed.) (1989) *Reflections on Chomsky*, Oxford: Blackwell Publishers.
Harman, G. (ed.) (1974) *On Noam Chomsky: Critical Essays*, Garden City: Anchor Books.
Hornstein, N. and Antony, L. (eds.) (forthcoming) *Chomsky and His Critics*, Oxford: Blackwell Publishers.
Kripke, S. (1982) *Wittgenstein on Rules and Private Language*, Cambridge, MA: Harvard University Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

35

Richard Rorty (1931–)

MICHAEL WILLIAMS

Richard Rorty has taught at Wellesley, Princeton, and the University of Virginia. Since retiring from Virginia, he has been a member of the Department of Comparative Literature at Stanford.

Early in his career, Rorty wrote extensively on topics in the philosophy of mind, emerging as an influential defender of eliminative materialism. But he was also concerned with metaphilosophical questions. His introduction to his anthology, *The Linguistic Turn*, surveys the history of the analytic movement with the aim of casting doubt on the view that, by centering philosophy on questions of language and meaning, analytic philosophy provides philosophers with new and more "scientific" methods for solving traditional philosophical problems. This argument foreshadows the radical turn taken by his mature work.

The main themes of this work emerge in a series of essays published in the 1970s and collected in *Consequences of Pragmatism* (1982). However, it was his book, *Philosophy and the Mirror of Nature* (1979), that made him the object of intense and often outraged critical scrutiny. In this book, he argues that philosophy, as practiced in mainstream Anglo-American philosophy departments, has exhausted its theoretical resources and outlived whatever usefulness it may once have had. It therefore deserves to come to an end.

Like other "therapeutic" philosophers, Rorty holds that our canonical "problems of philosophy" are to be avoided rather than solved. However, this is not because he sees them as pseudo-problems, rooted in misunderstandings or misuses of language. Rorty's approach is historicist. He denies that philosophy deals with perennial problems, intelligible to any reflective person because part of the human condition. Our canonical problems, then, are genuine enough, but only in the context of a historically contingent, hence potentially optional, configuration of ideas. Although they may once have promised great things, these ideas can now responsibly be dropped.

In Rorty's narrative, modern philosophy takes the form of epistemology or the theory of knowledge. Philosophy of this kind originates in the seventeenth century and achieves its definitive form in the writings of Kant. Descartes inaugurates modern philosophy's epistemological turn by making two moves: introducing methodological skepticism as the principal tool for investigating the foundations of knowledge, and redefining "mind" as that to which each of us has privileged access. Given this

conception of mind, skepticism itself acquires a new and more radical form. For the ancients, skepticism raised the question of whether we can attain certainty about the "real nature" of things. After Descartes, it raises the question of to what extent, if any, our "ideas" are accurate representations of "external" reality. The very existence of the external world is subject to doubt.

Descartes's philosophical project is foundational in two senses. It aims at identifying both epistemological foundations (certainties that resist skeptical challenge) and metaphysical foundations (the most basic explanatory commitments of the New Science). As Kant saw, the metaphysical aspirations of Descartes and his rationalist successors are problematic. Metaphysicians want to determine a priori, on the basis of our ideas alone, fundamental facts about the world. This cannot be done: Descartes's skeptical problem thwarts his metaphysical ambitions. Rationalist metaphysics is thus mere dogmatism.

Locke takes an important step towards a more purely epistemological conception of the philosopher's task by suggesting that, by investigating the powers of the Cartesian mind, we can determine the scope and limits of human knowledge. Locke, however, is insensitive to the powerful and general skeptical problem formulated by Descartes. Locke claims to investigate the limits of human knowledge. But, as Kant charges, in adopting a "historical" – i.e. empirical-psychological – approach to the origins of our beliefs, Locke fails to address the *epistemological* question of our *right* to hold them. Moreover, where the metaphysicians at least attempt to justify the basic presuppositions of modern science, Locke simply takes for granted the corpuscular-mechanical picture of the world.

Kant presents his transcendental idealism as the way beyond rationalist dogmatism and empiricist naturalism. His thought is that, since all empirically knowable objects, "outer" as well as "inner," are subject to conditions inherent in our cognitive constitution, we can have a priori knowledge of features necessarily characteristic of the world as we are able to know it. However, not all matters of human concern answer to these conditions of objective knowability. Those that do not remain matters of judgment or faith. Kant thus presents us with the idea of epistemology as a non-empirical discipline that determines the cognitive status of all other subjects according to how far they are controlled by reason and evidence, hence whether they aim at objective truth. Thus in modern philosophy, "refuting the skeptic," now conceived as establishing our right to claim knowledge of an objective, causally ordered world, ceases to be "the languid academic exercise of composing a reply to Sextus Empiricus" (Rorty 1979: 223), becoming instead the key to distinguishing between forms of discourse that are "rational," "scientific," or "cognitively significant" and those that are "emotive" or "merely expressive." Philosophy-as-epistemology becomes central to culture.

Michael Dummett argues that Frege, the founder of analytic philosophy, is as much a revolutionary as Descartes (see DUMMETT). In Dummett's view, Frege's revolution replaces epistemology, as the foundation of philosophy, with philosophy of language or "the theory of meaning," with the result that analytic philosophy is sharply discontinuous with philosophy-as-epistemology. Rorty sees no such discontinuity. Frege is a (notably original) member of the "back to Kant" movement. His turn to logic and language is an attempt to eliminate the Kantian tradition's last vestiges of psychologism, thereby rescuing philosophy from the scientific naturalism that was threatening to

overwhelm it. Analytic philosophy thus continues to pursue, in the idiom of "language," the epistemological questions that Kant and his predecessors pursued in the idiom of "ideas": segregating the cognitively significant from the merely expressive, drawing lines between the a priori and the empirical, showing where we should and should not be "realists" about truth, and so on.

A distinction that is absolutely essential to this Kantian style of philosophizing is that between scheme and content. Accepting this distinction, we will see empirical knowledge as involving two clearly distinguishable components, concepts and intuitions, or as resulting from the cooperation of two faculties, understanding and sensibility. On this model, "mind" or "language" orders or interprets the factual elements "given" to consciousness. Taken together, Rorty argues, Sellars's attack on "the myth of the given," Quine's skepticism about the analytic/synthetic distinction, Wittgenstein's critique of ostensive definition and "private language," and Austin's sarcasm about "the ontology of the sensible manifold" leave this fundamental commitment no longer credible (see AUSTIN, QUINE, SELLARS, and WITTGENSTEIN).

Rorty sees these critics of the Kantian tradition as united by a kind of methodological behaviorism. In their different ways, they invite us, first, to look at how we actually use words, revise beliefs, evaluate theories, or conduct inquiries and, second, to ask whether there is any payoff, theoretical or practical, in partitioning our beliefs or statements into "true-by-virtue-of meaning-alone versus true-by-virtue-of-fact" or "purely observational versus theory-laden." The answer is "No." The advantage of taking the linguistic turn, then, is not that it offers new ways of solving old problems but that it makes this methodological orientation plausible, thereby allowing us to set the old problems aside. In this way, analytic philosophy transcends and cancels itself.

The picture of inquiry and justification that results from abandoning the dualism of scheme and content is holistic, coherentist, and pragmatic. Inquiry is a process of constantly reweaving our web of belief under the impact of observation and in the light of multiple interests and criteria, theoretical and practical. Rorty thinks that this holistic picture blurs all the methodological distinctions – between the a priori and the a posteriori, the necessary and the contingent, fact and value, the sciences and the humanities, and so on – that philosophers bent on projects of epistemological or metaphysical demarcation want to keep alive.

Present in *Philosophy and the Mirror of Nature*, but much more strongly emphasized in subsequent writings, is the claim that the most fundamental error of our philosophical tradition is the notion that truth is correspondence with reality or accuracy of representation. The quest for truth-as-correspondence reflects an urge to be guided by something greater than ourselves: the World, the True, or the Good. (Rorty thinks of today's hard-headed scientific realism as evincing an essentially *religious* attitude.) This quest (which is as old as philosophy itself, philosophy-as-epistemology being simply its modern incarnation) is always associated with demarcational projects dividing matters of human concern into an upper and lower division: knowledge versus opinion, nature versus convention, philosophy versus poetry. However, in addition to undermining methodological grounds for such demarcations, the holistic, broadly coherentist and pragmatic conception of inquiry common to Quine, Sellars, and Wittgenstein makes it difficult to see individual sentences or beliefs as "corresponding" to anything. Whether

we look at inquiry from the standpoint of method or that of truth, we find no room for philosophy.

Rorty's focus on truth reflects an increasing self-identification with pragmatism. Having adopted a broadly coherentist picture of justification and inquiry, Rorty flirted briefly with the Peircean suggestion that truth is ideal justification. However, his settled outlook – which he identifies with the pragmatism of James and Dewey – is a radical anti-essentialism with respect to the traditional objects of philosophical concern. Rorty's Pragmatist does not replace a correspondence theory of truth with an epistemic account but rather holds that truth (or rationality or goodness) is not the sort of thing that we can usefully theorize about.

Rorty thinks that, among contemporary philosophers, Donald Davidson has done most to advance the pragmatist cause. According to Rorty, Davidson's work not only reinforces Sellars's rejection of "given" facts and Quine's repudiation of the analytic/synthetic distinction, it traces the connections between belief, truth, and meaning in a way that deprives these notions of all demarcational significance. For all their criticisms of traditional epistemology, Sellars and Quine are prone to backsliding because they remain committed to the view that the natural sciences, especially physics, get at "hard facts" or "the ultimate nature of reality" in a way that the softer disciplines do not. Davidson is able to go beyond Sellars and Quine because he is wholly free of this lingering scientism.

Perhaps because neither approach to truth makes our understanding of truth the key to traditional epistemological or metaphysical problems, Rorty pays scant attention to the distinction between Davidson's view that the concept of truth, while of considerable explanatory significance in the theory of meaning, must be taken as primitive, and the "deflationary" view that truth-talk is only an expressive convenience. Indeed, he often treats Davidson's view as a form of deflationism, a suggestion that Davidson emphatically (though perhaps not entirely convincingly) repudiates.

Another notable influence on Rorty's version of pragmatism is Thomas Kuhn. Rorty thinks that Kuhn's distinction between "normal" and "revolutionary" science invites wide application. In all areas of discourse, there are times when inquiry proceeds more or less normally, solving in agreed-upon ways commonly recognized problems, formulated in a familiar vocabulary. But sometimes we can make progress only by dropping old questions in favor of new ones, or by changing the basic vocabulary in terms of which our problems and projects are described. Rorty thinks that his own pragmatist attack on traditional philosophy is an instance of just such an attempt at revolutionary change.

Rorty's rejection of the correspondence or "realist" conception of truth is often thought to amount to an extreme form of linguistic idealism. If our beliefs do not answer to the world, truth is something we make up: the idea of objective truth goes by the board. Rorty thinks that the idea of "answering to the world" confuses causation with justification. Because we are trained in observation-reporting practices involving the causal triggering of reporting dispositions by external circumstances, the world plays a causal role in regulating our beliefs. But it does not play a justifying role. The situations that provoke such reports do not demand to be described in any particular vocabulary and do not determine the inferential or theoretical significance of the reports they provoke.

Critics sometimes charge that giving up on a substantive notion of truth, whether realist or Peircean, prevents Rorty from seeing inquiry as progressing. Rorty meets this charge by saying that improvements are measured retrospectively and comparatively – by reference to problems solved, improvements made, or alternatives foregone – rather than by their shortening the distance between ourselves and the End of Inquiry. We have no conception of what it would be for inquiry to have an end, no idea of "the Truth" as the Ideal Theory of Everything or the way that Nature itself would like to be described.

Rorty has also been widely criticized for preaching irrationalism and relativism. He rejects both charges. He agrees that his relaxed version of coherentism entails that justification is less algorithmic than many epistemologists have wanted it to be but denies that this is equivalent to the claim that anyone can (rationally) think whatever he likes or that any system of beliefs is as good as any other. Our settled beliefs, involuntary observations, and theoretical and practical interests provide all the constraint we need (and can possibly have). His position, he concedes, is "ethnocentric" in the following sense: at any stage of inquiry, we can only work with whatever beliefs and theories and criteria we have on hand. That is, we have to accept the irreducible contingency of our investigative and argumentative resources. Given this contingency, there are likely to be issues with respect to which, at any given time, not all people can find common ground. But this does not mean that some (or any) disputes reflect commitments that are in principle "incommensurable." We cannot predict the future of inquiry and never know how the dialectical situation will evolve. Rorty thinks that only disappointed foundationalists will equate his thoroughgoing fallibilism with skepticism, relativism, or irrationalism.

In recent years, Rorty's writings have taken a political turn. He defends a position he sometimes calls "postmodern, bourgeois liberalism": "bourgeois liberalism" because it fully endorses the rights and freedoms typically guaranteed by the rich, industrial democracies; and "postmodern" because it eschews the need for providing those rights and freedoms with a philosophical justification. Rorty recognizes that many philosophers think that, if we give up on such Enlightenment conceptions as universal reason and the Rights of Man – the kinds of thing philosophy is invoked to underwrite – we leave ourselves with no way of showing what is wrong with oppressive, discriminatory, or tribalist forms of political life. Indeed, he thinks that concerns about relativism and irrationalism grow out of just this fear. In reply, he advocates facing up to the "priority of democracy to philosophy." Democratic constitutions and the rule of law are appealing to people with our history and cultural background, but often to other people too, if they get the chance to enjoy them. Those who want philosophical foundations for liberal-democratic institutions should recall that such institutions did not appear overnight. Extending political rights and legal protections to all citizens, without regard to religion, race, or gender took time; and, in Rorty's view, this increasing inclusiveness owes more to an enlargement of sympathies than discoveries to the effect that rationality or moral considerability is more widespread than used to be thought. Imaginative literature and investigative journalism have done more for the oppressed and excluded than inquiries into the "foundations" of morals and politics.

Unusually for an American philosopher, Rorty has written extensively about such "continental" figures as Husserl, Heidegger, Foucault, and Derrida. He sees continen-

tal and analytic philosophy as having followed parallel courses. Like Frege, Husserl wanted philosophy to be rigorous and scientific, yet deeper than and prior to the special sciences. Also like Frege, he sought this depth and priority in a general account of representation. Unlike Frege, who turned to logic and language, Husserl looked for a theory of the invariant structures of consciousness. But he too provoked a pragmatist reaction. Roughly speaking, Heidegger (especially the Heidegger of *Being and Time*) stands to Husserl as the later Wittgenstein stands to Frege and Russell. The consequences of this reaction are further worked out in Derrida's deconstructive readings of seminal philosophical texts and Foucault's historicist reconstructions of vanished conceptions of scientific knowledge.

While he is generally regarded as arguing for the death of philosophy, this is a description Rorty repudiates. Following Sellars, he suggests that "philosophy" can be understood two ways. On the one hand, there is philosophy (little p): the attempt "to see how things, in the broadest possible sense of the term, hang together, in the broadest possible sense of the term." This Hegelian project of grasping one's time in thought could only come to an end if inquiry itself (in a broad sense that encompasses science, the humanities, literature, politics, and the arts) ground to a halt. On the other hand, there is Philosophy (big p): the Platonic–Kantian project of determining how to seek truth (or conduct oneself rationally to do more good) through discovering the nature of truth (or rationality or goodness). Where philosophy seeks reflective self-understanding, and perhaps self-transformation, but always at a particular stage of inquiry, Philosophy tries to discern the permanent framework within which all inquiry proceeds. In trying to kill off Philosophy, Rorty looks forward to a "post-Philosophical culture," in which such a quest will look as quaint as medieval theological disputes look to secular intellectuals today. Learning to do without Philosophy, as most intellectuals have learned to do without religion, means coming finally to take full responsibility for our opinions and values. Rorty's philosophy is thus a version of "humanism," in Sartre's sense.

Bibliography

Works by Rorty

1967: *The Linguistic Turn*, Chicago: University of Chicago Press.
1979: *Philosophy and the Mirror of Nature*, Princeton, NJ: Princeton University Press.
1982: *Consequences of Pragmatism*, Minneapolis: University of Minnesota Press.
1988: *Contingency, Irony, and Solidarity*, Cambridge: Cambridge University Press.

Work by other authors

Malachowski, A. (ed.) (1990) *Reading Rorty*, Oxford: Blackwell Publishers.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

36

John R. Searle (1932–)

A. P. MARTINICH

J. L. Austin was being recruited by the University of California at Berkeley in the late 1950s. He declined, saying, "I think I should be dead by then" and thereupon added, "Since you can't get me, get Searle." Searle became an assistant professor there in 1959, the same year he received his D.Phil. from Oxford. Aside from visiting appointments and leaves, he has spent his entire career at Berkeley, where he is Mills Professor of Philosophy.

In addition to Austin, Searle had been a student of P. F. Strawson, Peter Geach (who directed his dissertation), and other distinguished Oxford philosophers during the heyday of ordinary-language philosophy. Adept at criticism, Searle is even more impressive as a constructive philosopher. Even in some of his early articles, which are ostensibly criticisms of others, his own positive theory is not far below the surface.

Language

In "Austin on Locutionary and Illocutionary Acts," Searle shows how Austin's original linguistic distinctions should be recast. The most important result is that paradigmatic cases of illocutionary acts should be understood as consisting of a force and a propositional content (see AUSTIN). Consider, for example, these sentences:

I state that Jones will be at the party.
I promise that Jones will be at the party.
I question whether Jones will be at the party.

It is obvious that all of these sentences have something in common. Each would be appropriately used to express the same content or proposition that Jones is at the party. (For simplicity's sake, the temporal element will be ignored.) In these sentences, the propositional content is expressed with a "that" clause; but some sentences express their content with gerundive phrases or infinitives:

I congratulate Jones for being at the party.
I order Jones to be at the party.

It is equally obvious that a standard use of each of the above sentences expresses their propositional content with a different "force," the force of a statement, promise, ques-

tion, congratulation, and order, respectively. So the structure of illocutionary acts can be represented as $F(p)$. The p corresponds to the propositional content, and the F indicates the "force" attached to the proposition. In effect, Searle moved Austin's rhetic acts from the category of locutionary act to a subpart of an illocutionary act.

In "What is a Speech Act?" (1965) and *Speech Acts* (1969), Searle used promising to illustrate the appropriate form of analysis for illocutionary acts. With slight modification, Searle's analysis was this:

In uttering a sentence *T*, a speaker *S* promises an addressee *H* to do an action *A* if and only if

1 Normal input and output conditions apply.
2 *S* expresses the proposition that *p* in the utterance of *T*.
3 In expressing that *p*, *S* predicates a future act *A* of *S*.
4 *H* would prefer *S*'s doing *A* to *S*'s not doing *A*, and *S* believes *H* would prefer *S*'s doing *A* to *S*'s not doing *A*.
5 It is not obvious to both *S* and *H* that *S* will do *A* in the normal course of events.
6 *S* intends to do *A*.
7 *S* intends that the utterance of *T* will place him under an obligation to do *A*.
8 *S* intends [*i* –*1*] to produce in *H* the knowledge *K* that the utterance of *T* is to count as placing *S* under an obligation to do *A*. *S* intends to produce *K* by means of the recognition of *i* –*1*, and he intends *i* –*1* to be recognized in virtue of (by means of) *S*'s knowledge of the meaning of *T*.
9 The semantical rules of the dialect spoken by *S* and *H* are such that *T* correctly and sincerely uttered *T* if and only if conditions (1)–(8) obtain.

Certain aspects of Searle's analysis might be fine-tuned. A better analysandum is: "In uttering a sentence *T*, a speaker *S* explicitly and non-defectively promises an addressee *H* to do an action *A*." Perhaps another preparatory condition is appropriate: "*S* is able to do *A*." Notwithstanding these and other possible adjustments, the form of Searle's analysis is powerful and easily adapted to analyze the full spectrum of illocutionary acts.

Another merit of this form of analysis is that categories of conditions easily emerge from them. Most importantly, conditions (2) and (3) express requirements for the propositional content of the act. Conditions (4) and (5) express preparatory conditions. (6) expresses a sincerity condition. (7) expresses an "essential condition," which conveys the aim or goal of the act.

In *Expression and Meaning*, Searle developed a taxonomy of speech acts: assertives (for example, statements and asseverations); directives (for example, commands and suggestions); commissives (for example, promises and vows); expressives (for example, apologies and congratulations); and declarations and assertive declaratives (for example, declarations and verdicts, respectively). Unlike Austin's taxonomy, which was founded on no principles and Zeno Vendler's, which was founded on syntactic principles, Searle's taxonomy is semantically based (and also closely related to the types of conditions already described). First of all, illocutionary acts are categorized on the basis of their point or purpose, as expressed by their essential conditions. Assertives aim at committing speakers to beliefs. Directives and commissives aim at committing someone to a course of action. Expressives aim at expressing a mental state, such as happiness

or sadness. Declarations and assertive declarations aim at bringing about some fact about the world; when the Chairman of the Olympic Committee says, "I hereby open the Games," the games are thereby opened.

Another dimension of categorization is word/world fit. Thus, to assert "The door is open" is to aim at getting the words to fit or correspond to the way the world is. If the door is not open, the deficiency is with the words. In contrast, to command, "You will open the door," is to aim at getting the world to fit the words. If the addressee does not open the door, the "deficiency" is with the world. Of course this worldly deficiency tends to make the addressee, not the world, culpable. Taking our lead from these two examples, we can say that assertives aim at having their words fit the way the world is, while directives and commissives aim at getting the world to fit the way the words say it is to be. Expressives do not have a direction of fit but presuppose some fact about the world. The use of "I congratulate you on winning" and "I apologize for stepping on your foot" presuppose a victory and an offense.

A third dimension of categorization, related to sincerity conditions, is the psychological state expressed in the illocutionary act. Assertives and assertive declarations express the speakers' beliefs. Directives express the speakers' wants and desires. Commissives express the speakers' intentions. Declarations do not express any psychological state; warranted by institutions, declarations do not need sincerity.

A fourth dimension of categorization, related to propositional content conditions, concerns what illocutionary acts can be about. Assertives have virtually no restriction on propositional content. Directives and commissives must be about future actions. Expressives must be about present or past actions or conditions. Declarations can be virtually about anything, although there are limits. One cannot fry an egg by uttering, "I hereby fry this egg."

Other dimensions, such as the intensity of illocutionary point (suggesting versus insisting) and the way the act relates to the rest of the discourse (objecting versus replying), while informative, are not crucial to the basic taxonomy. As revealing as the taxonomy is, it will acquire even greater importance because of its connection with intentionality.

So far, our discussion has focused on the nature of illocutionary force. This focus suggests the novelty of speech act theory, for philosophers had concentrated on propositions for two and a half centuries. Let's now consider how Searle treats this hoary matter.

Paradigmatic propositions consist of a reference and a predication. Traditionally, reference is considered the basic way in which words relate to the world. The spirit of that tradition is captured by Searle's formulation of the axiom of existence namely, that everything referred to must exist. This invites the question of how the word or phrase used to refer gets hooked up to the world. The complete story requires a theory of intentionality (to be presented later) since reference depends on physical expressions such as words being backed up by inherently representational mental states. At this point, however, Searle is able to say that reference depends on "the axiom of identification," namely, that a speaker must be able to identify the intended referent for his audience in a nontrivial way. The object is identified either by the descriptive content of the referring expression ("the first human to set foot on the moon"), by the referring expression plus the context ("this one here"), or by a combination of the two ("this red shoe here") (Searle 1969: 80).

Traditionally, proper names have been the paradigmatic kind of referring expression; and the proper understanding of them has occupied philosophers for two millennia. To restrict our discussion to the last half-century, two theories have dominated: the descriptive theory and the causal theory. According to the causal theory, which Searle rejects, a proper name gets connected with its referent in virtue of a causal connection between that name, the speaker's intention to have the addressee identify a particular object through her use of that name, and that particular object.

In contrast, according to the descriptive theory, names refer to referents in virtue of their descriptive content. Taking off from Frege's views, Searle maintains that proper names have both sense (*Sinn*) and reference (*Bedeutung*), and the reference is a function of the sense (see FREGE). Consider a name like "Aristotle," which refers to Aristotle. That reference occurs in virtue of a certain descriptive content. Unlike general words like "red" or "human," which seem to be obviously tied semantically to redness and humanity, respectively, proper names seem to have a looser but nonetheless indispensable sense. The meaning of "Aristotle" is "the logical sum [inclusive disjunction] of the properties commonly attributed to him" (Searle 1969: 173).

Although Searle has never taken back any part of this view of the meaning of proper names, his focus seems to change in his later work. While insisting that some representational content must accompany the use of a proper name, he does not assert that the name's meaning is that representational content. Further, like the causal theorist, Searle maintains that the representational content is causally related to the external world (Searle 1983: 238). I think that Searle ought to abandon the claim that proper names have *Sinne*. From the fact that every use of a proper name must be accompanied by some representational content, it does not follow that that content is the meaning of that name. It further does not follow that there is any stable content, shared by the people who use that name. Searle may be conflating Frege's mode of presentation with the propositional content of an utterance (see Searle 1983: 249, 251). What is most important in Searle's theory, it seems to me, is the role of intentionality in reference. He thinks that this commits him to descriptivism because he takes his main opponents, Kripke and Donnellan, to discount intentionality. I believe that Searle's theory captures the intuitions of both descriptivism and the causal theory, without falling into the errors of either, and that he gives the wrong impression in claiming to be a descriptivist.

Concerning the other part of a proposition, predication, Searle accepts part of the asymmetry thesis of Gottlob Frege and P. F. Strawson. While reference is the act of picking out or identifying an object for the purpose of classifying or categorizing it, predication is the act of assigning a property to the referent. A subject-predicate proposition is true if and only if the referent has the property ascribed to it. For Searle, the distinction between subject and predicate or reference and predication is primarily one of function, that is, of how something is operating and only secondarily one of ontology. As an ontological issue, there were two basic choices: nominalism (the denial that properties, or universals, exist) or realism (the assertion that universals exist as much as individual material objects do). Searle's basic position is that there is no substantive issue here. Universals exist solely because predicates allow nominalization. Because of sentences like "Socrates is wise," we can form sentences like "Wisdom is a virtue." Consequently, although there are universals, they depend "merely on the meaning

of words" (Searle 1969: 105). Predicating a universal means using "a predicate expression in the performance of a successful illocutionary act" (Searle 1969: 121; see also 124). According to the traditional taxonomy, Searle would count as a conceptualist, a nominalist of a liberal expression; words are predicated of the objects referred to (Searle 1969: 124) and universals exist.

Consciousness

Since a speech act is a kind of human action that requires a mental representation of the world, a complete theory of speech acts will be part of a theory of mind. For most philosophers the main problem in this area is "the mind–body problem." What is the mind; what is the body; and how do they interact? For Searle, these questions are not problematic:

> The solution has been available to any educated person since serious work began on the brain nearly a century ago, and, in a sense, we all know it to be true. Here it is: Mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain. (Searle 1992: 1; see also 1984: 14–15)

Searle calls his view "biological naturalism," and says, "Mental events and processes are as much part of our biological natural history as digestion, mitosis, meiosis, or enzyme secretion. . . . [Further,] intentional states stand in *causal* relations to the neurophysiological (as well as, of course, standing in causal relations to other Intentional states), and . . . that Intentional states are realized in the neurophysiology of the brain" (Searle 1983: 1, 15; see also p. 90 and 1984: 21–2).

He is willing to admit that mental states causally supervene on brain states, but he is uneasy about the admission because supervenience was originally a logical relation and he worries that his causal form may be conflated with the logical one (Searle 1992: 124–6). Theories of supervenience are typically reductionistic. Is Searle's? Yes, but in only one of the various senses of reduction. Theoretical reductionism, for example, tries to express or reduce all the laws of one theory T_1 to those of another T_2. Once this is accomplished, it is standard to claim that the entities referred to in T_1 do not really exist and that only the entities in T_2 do. So theoretical reductions, like most reductions, ultimately aim at some ontological reduction. But Searle's does not. Rather, he explains that he espouses a form of causal reductionism because he holds that consciousness is causally reducible to the brain processes (Searle 1992: 116). This causal reduction, unlike most, does not have a corresponding ontological reduction because of the irreducible first-person ontology of consciousness. While a person's pain is no doubt caused by a certain pattern of neuronal firings in the thalamus and other parts of the brain, a complete specification of this pattern would still leave out "essential features of the pain" (Searle 1992: 117), namely, how the pain feels. When heat was reduced to mean kinetic energy and colors were reduced to the refraction of photons, these two entities were redefined in order to eliminate the subjective element in the perception of them. But a similar kind of redefinition of consciousness is not possible because there is nothing to consciousness except the subjectivity (Searle 1992: 121–3).

Searle's position is designed to avoid both materialism and dualism. Unlike a materialist, he affirms the existence of irreducible mental phenomena and denies that they are identical with brain states. Unlike a dualist, he asserts that mental properties are physical properties. Since many critics think that Searle is a property dualist *malgré lui*, something should be said about this. A property dualist holds (1) that there are only two kinds of properties, and (2) that all properties are mental or physical. Searle denies both (1) and (2). Either there is only one kind of properties, physical ones, or there are many kinds. And, if there are many kinds, none of them are mental or physical in the sense intended in (2). Searle thinks that dualism and materialism share the assumptions that give rise to the "mind–body problem." But there is no such problem any more than that there is a "stomach-digestion problem" (Searle 1992: 15; see also 1984: 14).

The serious issue in the philosophy of mind, in Searle's opinion, is the nature and structure of consciousness. He dealt with its nature in *The Rediscovery of the Mind* (1992), and with its logical structure in *Intentionality* (1983). Searle came to appreciate how the structures of intentionality have to be understood as ways that consciousness exercises itself after he wrote *Intentionality*. So it is sensible for us to begin with his views about consciousness.

Concerning the nature of consciousness, Searle has a negative and a positive project. The negative project is to show that current work on the nature of the mental, especially in cognitive science, is conceptually confused. The confusion does not just interfere with constructing a correct theory, but also motivates misconceived research strategies and covers over the fact that consciousness is the central phenomenon of the mind. His refutation centers on the "Chinese room" thought experiment. A person with no knowledge of Chinese is locked in a room. He is given a batch of Chinese writing (input); he has a rule book, written in his own language, that correlates the input with other Chinese writing and explains how to select or produce Chinese writing (output) that under certain conditions would be an appropriate follow-up to the input. To an observer outside the room and ignorant of the process *in camera*, it may appear that the person or mechanism inside the room knows Chinese. Of course, neither that person nor any conjunction of that person and anything else relevant to the outputting knows Chinese. He may not even know that he is dealing with Chinese or any language at all. He is performing totally formal operations, that is, the person treats the writing solely in virtue of physical shape (or other physical properties). The writing means nothing to the person.

Since the person in the room is doing just what a computer does, Searle concludes that computers do not have minds nor cognitive states (Searle 1992: 45).

Searle has variously described what the person in the room (and similarly any existing computer) lacks: a semantics, or an appropriate causal connection with the input and output, or understanding (see Searle 1992: 69). However, what seems to be most basic for Searle is the fact that neither the man in the room nor any existing computer has any understanding of Chinese. This is crucial because understanding requires consciousness and consciousness employs a unique kind of causation. This type, explained in *Intentionality*, may be described tentatively here as intentional causation (see Searle 1992: 107–9). The operation of intentional causation is important because some cognitive scientists have tried to circumvent the consequence of Searle's scenario by constructing a "room" or robot that can perform complex

operations. Suppose a robot is outfitted with television cameras, connected to levers and pulleys, powered by a motor that drives the robot on wheels to various locations to arrange and rearrange boxes or other objects. These scientists think that by increasing the complexity of the internal mechanisms and by having the robot be affected by and to affect its environment, they have undermined the Chinese room argument (Searle 1984: 34–5, 40–1).

The scientists are mistaken in thinking that complexity or generic causality is the issue. Rather, it is the nature of the controlling mechanism (consciousness) and the type of causation (intentional) that are crucial for Searle. In *Rediscovery of the Mind*, he argues that the fact that cognitive scientists think that the brain is a digital computer and that a digital computer can be constructed out of an infinity of materials, including paper or magnetic tapes, cogs and levers, "a hydraulic system through which water flows . . . an elaborate system of cats and mice and cheese . . . [and] pigeons trained to peck as a Turing machine" – all of these are actual examples from theorists – proves that the theory is bankrupt and irrelevant to the nature of the brain and mind (Searle 1992: 206). As he says, "we wanted to know how the brain works," and it is no answer to say that the brain works like any digital computer might work, where "digital computer" is defined so broadly that "stomach, liver, heart, solar system, and the state of Kansas" count as digital computers (1992: 208, and 1984: 36).

Another way of getting at the same general point is to say that essential to the operation of the brain is "intrinsic intentionality." While this concept will be explicated later, it can be understood provisionally as meaning that the brain causes states that are inherently representational. In contrast, no current computer and no computer to be designed in the forseeable future has intrinsic intentionality. For the sake of simplicity, let's say that the operation of every computer can be understood as involving the production of sequences of "1"s and "0"s. It is also common to think of "1" and "0" as syntactic (purely formal) entities, and one and zero as semantic entities, that is, the meanings of the syntactic entities. Even when "1" and "0" are not understood to be semantic entities, it is common to think that the computer itself takes those marks to refer to one and zero.

Searle objects to this view. Computers understand nothing because they do not represent anything to themselves. Rather, the computer designers and the computer users interpret the marks, "1"s and "0"s, in some way that is useful to them, some computational way. But the interpretation and the understanding are in the designers and users only and not at all in the computer. This has the consequence that the computer "1"s and "0"s are not even intrinsically syntactic entities, since syntax and semantics are correlative ideas (Searle 1992: 209–10). Syntax is not in physics.

It is no good to urge that computers can or do have physical states other than "1"s and "0"s. Searle's point still holds:

> *notions such as computation, algorithm, and program do not name intrinsic physical features of [computer] systems.* Computational states are not *discovered within* the physics, they are *assigned to* the physics. (Searle 1992: 210)

Searle's scenario of the Chinese room is powerful because the only plausible locus of the intentional causation required for understanding Chinese is in the consciousness

of the person in the room, and it is obvious that that person is not conscious of and does not know Chinese.

One consequence of Searle's criticism of standard cognitive science is that an entire level of theory disappears; the idea of "an unconscious mental process" and hence the idea of their principles loses all justification (Searle 1992: 239–40). There are then only two legitimate objects of cognitive studies, the brain and consciousness. The study of the brain belongs to neurophysiology; the study of consciousness belongs to philosophy and a still misguided cognitive science.

This brings us to Searle's positive project about the mind, namely, to give a non-reductionistic account of consciousness that is consonant with neuroscience. What is ontologically essential to consciousness according to Searle is subjectivity: "the mental is essentially a first-person ontology" (Searle 1992: 70; see also pp. 77, 94–5). It is this fact that makes the "first-person" understanding of it fundamental. Third-person access to consciousness via observation of behavior is inherently incomplete insofar as it has no access to the experience of consciousness. It is also derivative because while consciousness is essential to causing behavior, consciousness is logically independent of behavior (1992: 69). There can be consciousness with no behavior.

Searle identifies a number of features of human consciousness. We shall divide them into seven categories.

(1) Consciousness manifests itself "in a strictly limited number of modalities" (Searle 1992: 128). In addition to the traditional five – seeing, touching, tasting, smelling, and hearing – there is the sense of balance, bodily sensations, which includes "proprioception," that is, the feeling of how one's body and parts of one's body is oriented, and the stream of thought.

(2) Consciousness is unified with respect to both temporal continuity of impressions and the spatial unity of various impressions. Yesterday, today, and tomorrow are all part of the same temporal system; here, there, and the other place are part of the same spatial system; and the two form a spatiotemporal system.

(3) Consciousness is a necessary condition for intentionality and typically is intentional, that is, directed at objects. All intentionality is aspectual. This is easiest to see in visual perception; things are always perceived from a point of view and as being things of a certain kind (1992: 133). A related aspect of intentionality is the fact that consciousness has a focus and this in turn gives rise to the difference between figure and ground in Gestalt psychology. Also, attention is directed to some contents of consciousness more than to others. The driver of a car may be paying more attention to his vacation plans than to his driving; yet both are simultaneously conscious.

(4) Consciousness has a "subjective feeling." There is a difference between human consciousness and what it is like to be a bat or a porpoise.

(5) Although it is not a special feeling, there is an air of familiarity about the objects that a person is conscious of. Even the unfamiliar is familiar in the sense intended here. A person walks into an office building and expects it to have elevators; the elevators are found in a fairly predictable location; they are easy to operate; and the door opens to a floor, which, though never seen before, has enough familiarity about it that the appropriate room is discovered. There is a sense in which people have knowl-

edge of the world in a way that is more general than any particular bit of knowledge about it. The world is not strange and mysterious. This fact is highlighted by surrealist artists with their melting watches and ever-ascending-and-descending staircases. There is another kind of familiarity with the world: people know generally where they are and what time it is, in relation to many other places and times. Searle calls this general spatial and temporal familiarity with the world "situatedness" (1992: 141) (cf. CHOMSKY). Related to (5) is (6).

(6) This is what Searle calls "overflow," the feature that has some specific perception or belief connect to other beliefs seemingly without end in some elaborate, not fully articulatable web: these east Texas trees are pines, like the pines of California, but not exactly; they flourish in wet areas not quite marsh, etc. Perhaps this feature is closely related to Searle's concepts of the Network and the Background.

(7) Clearly many states of consciousness are suffused with a mood (elation, depression, cheer) even though a mood "never constitutes the whole content of a conscious state" (1992: 140). Construed broadly enough, every state of consciousness has some mood or other. For most people, there is a permanent low level of pleasure connected with consciousness, and for some, there is a permanent low level of displeasure. (See also 1998: 73–80.)

A theory of consciousness would not be complete without a theory of the unconscious. Searle's main thesis is that "every unconscious intentional state is at least potentially conscious" (1992: 132; see also p. 152, and 1984: 43–4). In holding this, he is opposing the idea, standard among cognitive scientists, of a deep unconscious that can never be made conscious. They purport to subtract consciousness from conscious mental states and to find the "computational mind" as the remainder. This is an attempt to effect a separation of intentionality from consciousness, after which a mistaken account of intentionality as computation is presented. Searle's proof of his thesis is roughly this: only mental states are intrinsically intentional in the sense described above. Unconscious mental states are intrinsically intentional. All intrinsically intentional states are aspectual. Unconscious mental states, as unconscious, exist only as neurophysiological events. Therefore, in order to be mental states at all, the unconscious ones must be capable of being brought into consciousness by the underlying neurophysiological events (1992: 155–9, 172).

Searle's theory of the unconscious permits a neat explanation for the existence of unconscious pains. For example, people with chronic back pain often awake from sleep with pain. It is plausible that the characteristic neurophysiological events that underlie the conscious pain are present during sleep, absent whatever additional events that bring the pain to consciousness. When these additional events get triggered, the pain becomes conscious and the sleeper awakes (1992: 164–7).

Further, consciousness is transparent to itself. When David Hume looked into himself, in addition to not finding the "mind" he was looking for, he did not find consciousness either because it is always directed at other things and cannot be an object of direct study (Searle 1992: 97). That is why nineteenth-century "introspectionism" in psychology was doomed to be a failure and why it is natural to think that consciousness is not part of the physical world. The transparency of consciousness also explains why it makes no good sense to say that each person has "privileged access"

to her own. That figure of speech requires one to think that a person enters a space separate from herself in which she alone can stand. But there is nothing in consciousness analogous to that space (1992: 97, 104–5, 170–1).

Intentionality

The cash value of Searle's theory of the nature of consciousness depends upon his ability to explicate the structures of consciousness. The most salient structures have intentionality; that is, being directed at something in the sense of representing something (Searle 1983: 3, 11–12). While not all conscious states are intentional – undirected anxiety and nervousness are not directed at anything, nor, I think, is pure joy – the most important forms are. In order to emphasize that intending to do something is only one kind of intentionality, Searle capitalizes the latter term, but I shall not follow that convention.

Because thoughts and feelings (prominent forms of intentionality) are expressed in language, Searle thinks that much of the structure of intentionality can be read off from the structure of speech acts. So, just as the basic form of an illocutionary act is $F(p)$, the basic form of an intentional State is $S(r)$, a psychological mode, such as believing, hoping, or loving, and a representational content. Like speech acts, the content is often a proposition, as in believing *that George Washington was the first president*. But it can also be an object as when Jones loves *Smith*. In this latter case, a distinction needs to be drawn between Jones's representational content, and the real person Smith, which is the intentional object of that content.

Four features shared by illocutionary acts and intentional phenomena are especially revealing.

(1) Analogous to the force and propositional content of speech acts, intentional states have a psychological mode and a content. "I state that you broke the vase" and "I order you to wash this floor" are structurally identical with "I believe that you broke the vase" and "I want you to wash this floor," respectively.

(2) Direction of fit: like statements, beliefs have a word-to-world direction of fit, and like orders, desires and intentions have a world-to-word direction of fit. Like apologies, sorrow has no direction of fit but presupposes something about the world, an injury caused by the agent.

(3) For illocutionary acts that have a sincerity condition, the expressed propositional content purports to be the content of an intentional state of the speaker. For example, stating that snow is white is representing that one believes that snow is white. And promising to go to the party is representing that one intends to go to the party. Further, the relevant intentional state is the sincerity condition of the illocutionary act. Both the sincerity condition and the intentional state of asserting that snow is white is the belief that snow is white.

(4) Both illocutionary acts and intentional states have conditions of satisfaction. A statement is satisfied if and only if it is true; an order is satisfied if and only if it is obeyed; a promise is satisfied if and only if it is kept. Correspondingly, a belief is satisfied if and only if it corresponds to the way things are; a desire is satisfied if and

only if it is fulfilled (idiomatically, desires are said to be "satisfied"), and an intention is satisfied if and only if it is carried out.

So far, nothing has been said about a traditional part of the concept of intentionality, namely, that what is intentional is always directed at an object. While Searle holds that some states have intentional objects, he denies that all of them do, because he does not hold the standard view about what an intentional object is. For him, it is always an existent object. The intentional object of believing that Jones is happy is Jones, the very person herself. The intentional object seems analogous to the referent of a reference. Both referents and intentional objects are idiomatically spoken of as being "in" mind, although neither, of course, could literally be spatially in the mind. And the referent is no more a part of the proposition expressed than the intentional object is part of the representational content of a mental state or event. Given the analogy between intentional objects and referents, one might think that it forms the basis for a fifth shared feature. But Searle does not go this route; and in general he mentions intentional objects only to put them aside.

In addition to these four parallels between illocutionary acts and intentional states, there is an important disanalogy. Intentional states are intrinsically intentional, while illocutionary acts are not. Illocutionary acts have to be realized in utterances (sounds, marks, or gestures) that are intrinsically physical and only derive their intentionality from the fact that humans intend them to represent states that are intrinsically intentional (Searle 1983: 27; 1992: 78–80). In other words, when a speaker utters a sentence, he (i) intends that sentence to have conditions of satisfaction that are given by (ii) the intentional state expressed. The parenthetical roman numerals indicate what Searle calls a "double level of Intentionality in the speech act," namely, (i) the intention of giving the utterance conditions of satisfaction that are (ii) the intentional states. The condition of satisfaction of "snow is white" is that snow is white. To ask for the meaning of a sentence is to ask for the intentional state that is the condition of its satisfaction (1983: 28, 164).

What determines that some intentional state or event is one with some particular content rather than some other? In *Intentionality*, Searle's answer was that an essential element is its place within a system of intentional states. For example, the intentional state of running for the presidency of the United States is possible only as part of a Network of other intentional states involving such beliefs as that there is a United States, constituted in the eighteenth century, separate and independent of every other country, bounded in large part by Canada on the north and Mexico in the south, and so on. So Searle endorses a kind of holism.

Further, the Network of intentional states is possible only against a background of nonrepresentational "mental capacities" (Searle 1983: 21). The background is generalized know-how. It is a worldly competence that makes possible particular kinds of know-how, such as knowing how to open a door or knowing how to write a letter. But, because particular know-hows can be explicated as essentially involving propositional contents, such as that something is the case or that something is to be done, these instances of know-how are not part of the background itself. Opening a door involves representations, but "the ability to recognize the door and the ability to open it are not themselves further representations" (1983: 143). It is difficult to be explicit and precise

about the background just because background capacities are not propositional and explicit while explanations are.

Searle may be discussing the same phenomenon as Martin Heidegger when he talked about human beings as beings-in-the-world with things being ready-at-hand and as Ludwig Wittgenstein when he talked about certain things being so basic that they "stand fast" and hence are objects neither of certainty nor doubt. In this spirit, Searle claims that metaphysical realism is not a "hypothesis, belief, or philosophical thesis . . . but the precondition" of any of these things (1983: 158–9). Consciousness is not a hypothesis either (1992: 79; see also 1995: 178, 195).

Searle's views about the Network and the Background changed in *The Rediscovery of the Mind* although the core remained. "Intentional states do not function autonomously," he says. Their conditions of satisfaction depend upon a set of Background capacities. Some of these capacities are capable of generating other conscious states. Of course these newly generated conscious states are just like the first ones mentioned; they do not function autonomously and they depend upon a set of Background capacities. The Background role is so important that the "same *type* of intentional content can determine different conditions of satisfaction" when the Background capacities are significantly different (1992: 190). Consciousness then consists of an occurrent representational content (item (a)), which depends upon a neurophysiological base with "the capacity to generate a lot of other conscious thoughts." This capacity (item (b)) itself is part of a neurophysiological system that is nonrepresentational but necessary for both the representational content and the individual consciousness. Item (a) replaces Searle's earlier idea of a Network; item (b) replaces the earlier idea of the Background (1992: 190–1; see also 1995: 181–2, 184–9).

Having laid out the general structure of intentionality, we can look at Searle's treatment of the "primary forms of Intentionality, perception and action." Concerning perception, he is a realist; people see cars, tables, trees, and so on. They do not see their perceptual experiences; they have them (1983: 36, 38). Experiences, like beliefs and desires, are intentional, being directed at objects and having conditions of satisfaction. The experience of perceiving a yellow car is directed at the car and is satisfied only if there is a yellow car where it is perceived to be. While the car is yellow and car-shaped, the experience is neither. Further, the car itself has to cause the visual experience. How should this information be represented? Searle proposes the following: "I have a visual experience (that there is a yellow station wagon there and that there is a yellow station wagon that is causing this visual experience)" (1983: 48).

The crucial point about this analysis of the conditions of satisfaction of visual perception and related phenomena is the element of self-referentiality: the visual experience that is to be satisfied is mentioned in the conditions of satisfaction. This does not mean that the self-referential causal relation is itself seen (Searle 1983: 48–9). Searle has some sympathy with the suggestion that a clearer way to represent the conditions of satisfaction is this: VE_c (that there is a yellow car), where the subscript "c" indicates the causal self-referentiality.

The visual experience is actual seeing only if the appropriate object in the nonmental world, in this case, the right yellow station wagon, is causing the experience. This allows Searle to say that on his account

> perception is an Intentional and causal transaction between mind and the world. The direction of fit is mind-to-world, the direction of causation is world-to-mind; and they are not independent, for fit is achieved only if the fit is caused by the other term of the relation of fitting, namely, the state of affairs perceived. (Searle 1983: 49; see also pp. 61–2)

Does Searle's account solve a familiar philosophical puzzle? Suppose that two identical twins have type-identical visual experiences while looking at two different but type-identical station wagons. What makes one twin's perception a perception of car A and the other twin's perception a perception of car B? Searle says that there must be something in the representational content itself that specifies the proper car as a condition of satisfaction. The specification is achieved because "each experience is self-referential" (Searle 1983: 50). The visual experience of one twin includes as a condition of satisfaction the fact that *that very experience is being caused* by a yellow station wagon. In other words, perceptions as a mode of consciousness are essentially first-person phenomena. In contrast, a causal theory of perception according to Searle takes a third-person perspective: car A causes one perception and car B causes the other. The theory is inadequate, however, as Searle observes, because it has no account of how the perceiver's intentionality matters to the perception (1983: 64).

While Searle is right in his criticism, he does not seem to concede as much to the causal theorist as he should. The fact that any perceiver has her own individual visual experience with a self-referential content depends on the fact that some individual non-intentional object is causing that unique experience. That is, the object of perception helps individuate the visual experience.

Searle's opposition to the causal theory inclines him to assert that human beings are "brains in vats," the vats consisting of their skulls (Searle 1983: 230). Although he may have primarily intended the then fashionable phrase "brains in vats" metaphorically, it is hard to see how this position is consistent with central features of his theory. A brain in a vat does not receive its intentional contents in the right kind of way; its prior intentions do not cause actions because the brain is not hooked up to the appropriate biological organs, and so a brain in a vat does not perform any non-mental physical action.

A different, and, I think, ineffective objection to Searle's theory of perception is that it leads to a familiar form of skepticism. (1) Since the car causes the visual experience, the visual experience is the basis for believing that the car is seen; and (2) since the perceiver infers from the perceptual experience that the car exists, it might be the case that the experience exists without the car existing. Searle rejects both parts of this line of reasoning. The perceptual experience is not evidence for the belief that the car is seen; and the perceiver does not infer that the car exists. One simply sees the car: "The knowledge that the car caused my visual experience derives from the knowledge that I see the car, and not conversely. . . . [W]e perceive only one thing and in so doing have a perceptual experience" (1983: 73, 74).

Let's now consider Searle's theory of action. Intentional actions are conditions of satisfaction of intentions to act; but not every intentional action is preceded by a prior intention to act. People often act without planning (1983: 84–5, 107). That is, "prior intentions" should not be confused with "intentions in action." Prior intentions are like plans; they temporally precede an action. They may be expressed by sentence-forms like, "I intend to *A*" or "I will *A*." In contrast, an intention in action is part of the warp and

woof of an action. All actions are intentional. So-called unintentional actions, such as Oedipus marrying his mother, is related to something intentional, namely, Oedipus marrying Jocasta. Breathing, snoring, and sneezing are bodily movements but not actions, because they are not intentional.

Like perception (and memory), both "prior intentions and intentions in action are causally self-referential" (Searle 1983: 85). Intrinsic to their conditions of satisfaction is a causal relationship between the intentional state and the thing done. To raise one's arm is to have one's intention of raising it cause it to go up (p. 86). Searle says that "the prior intention causes the intention in action" (p. 94). This is true whenever there is a prior intention to do something immediately; such intentions cause at least attempts or efforts. To try to do something is to do something. However, prior intentions to do some act *A* are not sufficient to cause the person to do *A*. There is a gap between the intention and even the decision to act and the acting itself.

Perception and action are nicely contrasted by two facts: (1) While perception has a mind-to-world direction of fit, action has a world-to-mind direction of fit. (2) While an object causes a perceptual experience, an experience of acting causes an event. But perception and action are the same insofar as both present, rather than represent, their experiences, in contrast with, say, memory or imagination, which do re-present things.

Searle's researches into the logical structure of intentionality caused him to rework the analysis of meaning that he gave in *Speech Acts*, where his main goal had been to give an analysis that would not include intending to cause an effect, as H. P. Grice had in his. In *Intentionality*, Searle claims that the essence of meaning is representing; in effect, it is imposing conditions of satisfaction on something that is not inherently representational. For example, in raising his arm, a person means that the army is retreating if and only if his intention to raise his arm causes his arm to go up and his arm raising has as a condition of satisfaction that the army is retreating. Obviously, the fact that the arm raising has a condition of satisfaction is due to the mind imposing that condition on it. All meaningful gestures or utterances have an intentionality that is derived from mental states that are inherently intentional.

Although representing is the heart of meaning, according to Searle, since meaning is standardly used to communicate, a complete account has to say how communication occurs. What needs to be added to meaning as representation is an intention to get the audience to recognize what the condition of satisfaction of the person's gesture or utterance is and to have the audience recognize it in virtue of the gesture or utterance itself (Searle 1983: 168).

One merit of this analysis is that it gives a precise and informative answer to the question "What is the difference between saying something and meaning it versus saying something and not meaning it?" The answer is that the former has conditions of satisfaction and the latter does not. When "Es regnet" is said and meant, a condition of satisfaction is that it is raining, whereas when it is said merely in the course of practicing German, the weather is irrelevant.

Social reality

At the end of *The Rediscovery of the Mind*, Searle offers some guidelines for the proper study of mind; the last of these, the last sentence of the book, is: "we need to rediscover

the social character of the mind" (1992: 248). This was an advertisement for his next book, *The Construction of Social Reality* (1995). The title is a direct challenge to a diametrically opposed idea, the social construction of reality. This popular alternative, which maintains that the world is a construct of the human mind, has no appeal for Searle, who says it is a "preposterous" view that rests on "an array of weak or even nonexistent arguments" (1995: 160). He is a realist about the real world.

The natural world consists of two basic kinds of things: non-mental and mental. The non-mental things at a relatively fundamental level are atoms and at the most fundamental level are space-time points, according to current science. The mental things evolved out of the non-mental after billions of years. The non-mental things are ontologically objective; they exist independently of minds. The mental things are ontologically subjective; they depend for their existence on minds. For example, a pain is a part of the natural world, but subjective.

The social world, which includes money, government, and marriage, arises out of mental reality, largely because the mind can represent things as other things. To take the most powerful instrument of representation, language uses sounds, marks, or gestures to represent other things; and this is possible because people are willing and able to take them as representing other things. In short, the social world is observer- and user-dependent. Money is money because people take pieces of paper or metal to have exchange value. Citizens are citizens because people treat them as having certain rights and responsibilities. In contrast, the natural world is observer-independent. A rose is a rose is a rose, whether anyone views it or not.

Given this ontology as background, Searle uses three elements to explain social reality. First is the idea that people can impose functions on objects that do not have that function beforehand. Never "intrinsic" to the thing itself, functions are always observer-dependent and introduce a normative dimension (Searle 1995: 19). Originally, some object became a hammer when it was used to hammer and not before; and, in becoming a hammer, it became possible to judge good and bad hammers, depending upon how well they functioned.

The second element is collective intentionality. A lineman blocks an opposing linebacker only as part of his team's play, and a musician plays first violin only as part of the orchestra's symphony. Football players play as much in concert as musicians do (Searle 1995: 22). Collective intentionality is not reducible to individual intentionality. In doing something together, each participant has her own individual intentions, but these derive from the collective intentionality of the group (pp. 24–5).

The third element is the distinction, introduced by Searle in *Speech Acts*, between regulative and constitutive rules. Regulative rules direct or control pre-existing behavior, as the rules of etiquette control how people should eat. They typically take the form of imperatives: Use a napkin, not your sleeve, to wipe your mouth; eat your peas with a fork, not your knife. Constitutive rules create new forms of action. The rules of football create the game of football. More importantly, the basic rules of a government, formulated in constitutions, create governments. When one government falls and another arises, a new constitution is formulated. These constitutive rules often take the form of indicatives: "The Supreme Court is the highest court of the judicial branch of the United States." Such sentences may seem to be statements of fact but they are more properly seen as declarations (Searle 1995: 55, 74). Searle believes that the deep form of con-

stitutive rules is "X counts as Y (in context C)," for example, "A person born on American soil counts as an American citizen" (p. 28). In keeping with the theory of *Speech Acts*, one might suggest that a deeper form is, "We declare that X is Y (in context C)" (cf. 1995: 104–11). The suggestion is tempered by the fact that so many institutional acts evolve slowly, haltingly, and unreflectively.

The "X counts as Y" formula can be iterated. Something that once occurred as a Y term, for example, "citizen," can occupy the place of the X term in another formula. "A citizen counts as the president when duly elected, etc." Roughly, the more complex the society, the more numerous and iterated the "counts-as" formulae (Searle 1995: 80).

When constitutive rules are enacted, institutional facts are created. These facts are self-referential: something is a five dollar bill because it is accepted as a five dollar bill. At a deeper level, a level that exploits the three elements of function, collective intentionality, and constitutive rules, the logical structure of institutional facts is this:

We collectively accept (S is enabled/required (S does A)).

Applying this structure to a five dollar bill, call it X, we get, "We accept (S, the bearer of X, is enabled (S buys with X up to the value of five dollars))" (Searle 1995: 97–8, 104–12). Contrary to appearances, this analysis is not viciously circular, because "five dollar bill," and, more generally, "money," occupy only a couple of nodes in "a whole network of practices, the practices of owning, buying, selling, earning, paying for services, paying off debts, etc." (p. 52). The range of institutional facts is enormous, from "wives to warfare, and from cocktail parties to Congress" (p. 96). They, and the powers that go with them, come into existence when people accept them as facts and continue to exist as long as people accept them as facts.

Conclusion

The claim that Searle counts as a philosopher of the first rank turns on this point: he uses a small number of interlocking elements to explain a broad spectrum of reality in an illuminating way. The most important elements are the ideas of representation, direction of fit, self-referential intentional causation, and the distinction between constitutive and regulative rules. The spectrum includes the nature of language, mind, and the social world, all presented within a naturalistic but not materialist world-view.

Bibliography

Works by Searle

1969: *Speech Acts*, Cambridge: Cambridge University Press.
1979: *Expression and Meaning*, Cambridge: Cambridge University Press.
1983: *Intentionality: An Essay in the Philosophy of Mind*, Cambridge: Cambridge University Press.
1984: *Minds, Brains and Science*, Cambridge, MA: Harvard University Press.
1985 (with Vanderveken, Daniel): *Foundations of Illocutionary Logic*, Cambridge: Cambridge University Press.

1992: *The Rediscovery of the Mind*, Cambridge: MIT Press.
1995: *The Construction of Social Reality*, New York: Free Press.
1998: *Mind, Language and Society*, New York: Basic Books.

Work by other authors

LePore, Ernest and Gulick, Robert van (1991) *John Searle and His Critics*, Oxford: Blackwell Publishers.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

37

Jerry Fodor (1935–)

GEORGES REY

Jerry Fodor is widely regarded as the most significant philosopher of mind in recent times. With Noam Chomsky at MIT in the 1960s he mounted a decisive attack on the behaviorism that then dominated psychology and most philosophy of mind, and has tried to present in its place a naturalistic and realist account of mental processes that renders them amenable to serious scientific study.

Indeed, he is one of the few philosophers who has combined philosophical and empirical psychological research, publishing work in both domains, developing at least two theories that have become highly influential in both of them: the computational/representational theory of thought processes (see the section "CRTT: Computation") and the modularity theory of perception ("Modularity and the Limits of CRTT"). These theories are, however, best appreciated against the backdrop of a number of other themes in Fodor's work, which provide the best overview of his work, as follows: (1) Intentional Realism; (2) Nomic Explanation; (3) The Problems of Mind; (4) CRTT: Computation; (5) CRTT: Representation; (6) Solipsism and Narrow Content; (7) Nativism; (8) Modularity and the Limits of CRTT.

Intentional realism

Fodor's primary concern is to defend the familiar "belief/desire," or "propositional attitude" psychology with which the folk routinely explain each other's behavior; for example, someone's heading south is explained by their wanting water and thinking there's some there. What we might call "scientific intentional realism" is simply a way of taking folk psychology seriously as the beginning of a serious scientific psychology.

Not that the folk are always right about the mind. Indeed, many of the claims and even the specific terms they employ: "learning," "memory," perhaps even "belief" and "desire," may well turn out to be theoretically inadequate. Fodor only presumes that, whatever the particular kinds of phenomena invoked by an ultimate psychology, they will display certain crucial properties:

1 as some or other species of *propositional attitude*, they will be *intentional* (being "about" things) and *semantically valuable* (capable of being *true or false*);
2 as phenomena involved in rational thought, they need to be *logically structured*;

3 as ultimately explanatory of action, they need to be *causally efficacious*; and
4 given the fundamental role of physics in our understanding of the world, they had better be *composed of or identical to* physical phenomena.

Fodor's postulation of states displaying these properties may seem rather truistic until one notes the quite substantial efforts of a good number of philosophers and psychologists in the twentieth century to argue otherwise: there are not only the usual Cartesian dualists, who resist the materiality of the mind (and hence (4)), but more recently there have been the logical positivists, behaviorists, Wittgenstein, Quine, Gibson, Davidson, Dennett, the Churchlands, many connectionists, all of whom have tried in one way or another either to deny the causal reality of the mind altogether, or to relegate mentalistic ways of describing the world to some sort of "second grade status," in some way less objective than physics (denying (3)). Much of Fodor's work consists in defending intentional realism against these attacks, not only as they arise in philosophy, but particularly in relation to psychology, where what is at stake are entire research programs committed to behaviorism, connectionism, neurophysiology – or, as he would recommend instead, to intentional realism (1968a, 1998b: chs 1, 8–10).

Fodor is specifically concerned with the kinds of challenges to a materialist mentalism that were raised by Descartes's concern with rationality, and Brentano's with intentionality, and so is consequently most concerned to be a realist both about attitude *states*, such as belief or desire, and about their *contents* (the belief that *snow is white* or *God is dead*). Indeed, he is adamant that the latter are decidedly not to be explained away as matters of "interpretation" or mere "similarity relations" (1998a: 30–4). He is also as realist as anyone about "qualia" and consciousness, but has relatively little to say about them. He is convinced that empirical psychology at least since Freud has given us reason to suppose that rational and intentional phenomena needn't be conscious (1968a, 1998b), and that he can therefore address the formidable difficulties of the former without worrying about what he regards as the currently impenetrable problems posed by the latter (1994: 121, and 1998b; but see 1972 and 1998b: 73 for some stray substantive remarks).

Explanation as nomic subsumption

Fodor also takes it for granted that explanation in general is subsumption under laws, and that the realm of the mental is no exception (1994: 3, 1998a: 7). Except possibly in ultimate physics, he assumes these are *ceteris paribus* laws, or laws that are not "strict" and "exceptionless," but hold in abstraction from various interferences or "completers" that a fuller theory of the world might include (1991c). Much of his view here is of a piece with the kind of idealization that Chomsky noted is typical of any science (see CHOMSKY). But whereas Chomsky is largely concerned with only a specific set of idealizations – those capturing linguistic "competence" in abstraction from its "performance" – Fodor is concerned with what he regards as the necessary variety of them that are enlisted in the explanation of psychological *processes*.

Towards addressing this concern, Fodor (1968a) presented one of the first lengthy defenses of *functionalism*, according to which psychological states are

individuated by their causal relations. Since different physical phenomena can satisfy these relations, functionalism naturally gives rise to cross-classificatory *layers* of explanation: one level of causal relations may be "multiply realized," or variously "implemented," by different mechanisms at a lower level (1968a, 1998b: ch. 2). Specifically, the *intentional level* of a cognitive psychology may be implemented at a lower level by various computational/syntactic processes (§4), which in turn may be implemented by different physical mechanisms – brains in the case of people, transistors in the case of machines. (Note, though, that Fodor is nevertheless skeptical one can provide any *definitions* of mental states, functionalist or otherwise; see §5.1).

Especially in Fodor's work, this functionalist conception is responsible for a considerable "autonomy" of cognitive psychology from details of its implementation, analogous to the way a computer program can be specified in abstraction from the electronic details of the computers that run it. Because they are genuine laws, involving, for example, "projectible" predicates, the kinds they mention are not reducible to mere finite disjunctions of the kinds at the lower levels (1998b: ch. 2), although Fodor presumes that they "supervene" on them.

The demand for mind

If we live in a purely physical universe, however, it might be wondered what serious explanatory role mental phenomena have to play. Why doesn't physics alone suffice? The short answer is, of course, that purely physical processes can come to exhibit all manner of special structures and organizations – molecules, crystals, cells, living organisms, and sometimes minds – that it is the business of special "macro"-sciences to describe.

Fodor is particularly impressed by the sensitivity of human beings to indefinitely many non-local, non-physical properties: not only, as Chomsky has emphasized, to highly *abstract* grammatical properties, like being a morpheme or a noun phrase, but also to arbitrary *non-physical* or *non-local* properties, such as *being a crumpled shirt*, *a grieving widow*, or *a collapsing star* (1986). These sensitivities are particularly impressive given that they seem to be (1) *productive* and (2) *systematic*. (1) People seem capable of discriminating stimuli of indefinite logical complexity, such as *being a crumpled shirt that was worn by the thief who stole the cat that chased the rat . . .* (1975a: ch. 1); and (2) anyone capable of discriminating one logical form is capable of discriminating logical permutations of it; for example, one can discriminate *John's loving Mary* if and only if one can discriminate *Mary's loving John* (1987b: 147ff.).

A good deal of Fodor's work has been devoted to showing that no non-mentalistic account can explain these phenomena. Thus he has argued at length that purely physicalist, behaviorist, Gibsonian, syntactic, and eliminative connectionist accounts of behavior are either vacuous or empirically inadequate (1968a, 1981b, 1987b: 161–3, 1988a, and 1991a). It is difficult to see how any physical mechanism could be sensitive to such an extraordinary range of arbitrary properties of the world without exploiting internal processes of logical combination, inference, and hypothesis confirmation that essentially involve phenomena satisfying the four demands listed on pp. 451–2.

CRTT: Computation

Fodor's main proposal for meeting those four demands is the *computational/representational theory of thought* (CRTT). Indeed, much of his work can be regarded as an effort to incorporate into psychology Alan Turing's crucial work on mechanical computation, according to which certain rational processes could be realized mechanically: for example, each of the rules of logic can be shown to involve mechanical operations on the formally specified sentences of a formal language. Fodor regards this as promising for psychology, since, he argues, people at least sometimes engage in the sort of *truth*-preserving inferential processes captured by logic (1994: 9). This already marks an important break with traditional psychology, which tended to rely on mere *associations* among ideas (Hume) or stimuli (Skinner). These seem incapable of capturing the relation between, for example, the premises and the conclusion of a valid argument; mere associations, like that between "salt" and "pepper," are neither necessary nor sufficient for understanding valid arguments.

This concern with logic and truth also commits psychology to vehicles capable of the requisite representational richness. Traditional empiricist psychology (as in Locke and Hume) tended to be not only associationist, but also to regard mental representations (or "ideas") as *images*. But although images may have a role to play in thought (1975a: 174–94), it is doubtful that they are remotely adequate for the expression of it in general. What image could express the conditionals, quantifiers, negations, and modals in such a thought as *If everyone drinks then no one should drive*? Merely a picture of a lot of non-driving drinkers won't quite do. The only vehicles that seem remotely capable of expressing such thoughts are the logico-syntactic vehicles of a language, natural or artificial, with precisely the resources of operators (quantifiers, connectives) and referential devices (predicates, variables, names) that we ordinarily use to express those thoughts. That is, there must be some sort of language in which a thinker thinks, a "language of thought" (an "LOT" or "mentalese").

Talk of "sentences" in the brain mustn't be taken on the model of sentences as they are inscribed on pages of books. Sentences are highly abstract objects that can be entokened in an endless variety of ways: as wave forms (in speech), as sequences of dots and dashes (Morse code), as sequences of electrically charged particles (on recording tape). It is presumably in something like the latter form that sentences would be entokened in the head. Indeed, CRTT is best viewed as simply the claim that the brain has logically structured, causally efficacious states, a thesis that, whatever its merits, isn't *patently* absurd. (Note also that this is not a thesis that is supposed to be *introspectibly* plausible: CRTT does not entail that people's mental lives should appear "introspectively" to involve sentences, much less sentences of a natural language.)

An extremely simple version of CRTT could be true of an intelligent system in the following way: there are sensory modules (e.g. visual and auditory systems, see "Modularity and the Limits of CRTT," below) that transduce ambient energy forms into electrical signals that in turn produce structured sentences as input to a central cognitive system (perception). This central system selects certain sentences from a pre-established (perhaps innate) set, tests their deductive consequences against this input

for a "best fit," and produces as output those sentences that pass that test above threshold (belief). These sentences in turn may be the input to a decision-making system in which, on the basis of that input, innate preferences, and utility functions, a course of action is determined, that is, a basic act-description is selected (intention) that causes a basic act satisfying that description to be performed (action).

A common objection to such an account is that it requires "homunculi" in the brain, possessing precisely the same intelligence that is supposed to be explained, in order to "read" these sentences. However, what Turing's theory of computation has taught us is that complex computational processes, such as operations upon symbols in a language, can be broken down into simpler operations that are eventually so simple that (it's obvious that) a machine can execute them (1975a: 73) without at that point exploiting any intelligence at all.

Besides the straightforward argument from the expressive power of a language, Fodor advances a number of other reasons for CRTT:

1 It is presupposed by standard theories of perception, hypothesis confirmation, and decision-making, all of which involve the agent representing the world and assigning utilities to the consequences of various courses of action (1975a: ch. 2).
2 It is able to explain the *truth*-preserving transitions in thought of which rational creatures are at least *sometimes* capable: for example, the ability to deduce "Rats die" from "Cats live" and "If cats live then rats die" (1994: 9).
3 It is easy to conceive of computational architectures exploiting an LOT that would explain the productivity and systematicity of our minds mentioned earlier, for example, a machine that was sensitive to syntactic structures could in standard recursive ways produce indefinitely complex representations (productivity), and could access one representation if and only if it could access a logico-syntactic permutation of it (systematicity) (1987b).
4 It offers a perspicuous account of the intensionality of thought: Oedipus' thinking he'll marry Jocasta is distinct from his thinking he'll marry his mother since the vehicles of the two thoughts (the LOT equivalents of "Jocasta" and "my mom") are different.

It is at least these four arguments that Fodor deploys both against traditional associationism in psychology, as well as against what he regards as its contemporary manifestation, radical connectionism (1988a, 1991a). (Against connectionist proposals that would merely propose a novel *implementation* of CRTT, Fodor has no objection.)

Despite his commitment to a CRTT, Fodor has doubts about its eventual scope, to which we'll return below (in "Modularity and the Limits of CRTT").

CRTT: Representation

So far, we've discussed Fodor's views about mental *processes*, as computations over logico-syntactic representations. Fodor would be the first to recognize that this is at best only half an account of mentality: computations may well preserve truth better than associations do, but where do these representations acquire any semantic properties like truth in the first place?

Inferential role theories

In early work (1963), Fodor was drawn to what is, broadly speaking, an "inferential role semantics" (IRS). This is a family of views according to which the meaning of an expression has to do with its inferential relations to other expressions, as in the case of "bachelor" entailing "unmarried." These relations might involve definitions (as in traditional "analyses" and "meaning postulates" in philosophy), "semantic decomposition" in linguistics, "procedural semantics" in artificial intelligence, or "prototypes" and "whole theories" in psychology.

By the late 1970s Fodor became convinced that the standard arguments for IRS suffered from serious empirical and philosophical difficulties: proposed linguistic decompositions were seriously inadequate; there was an embarrassing paucity of psychological evidence for anything like definitions (1975b), which the history of analytic philosophy had shown were notoriously difficult to provide in the first place (1970); and Quine had cast serious doubt on whether there could ever be any theoretically satisfactory way of distinguishing constitutive inferential relations (the "analytic") from merely common beliefs (the "synthetic") (1998a) (see QUINE). Indeed, although Fodor has no patience with Quine's behaviorism, he wholeheartedly endorses his rejection of definitions as having any serious explanatory status in any science whatsoever (the intuitive appearance of an "analytic/synthetic" distinction, Fodor argues, is due either to the "centrality" of a claim to one's thought, or to its involving "one-criterion" concepts (1998a: 80–6)).

Under the influence of Quine, and especially of his dictum, "The unit of meaning is the whole of science," many IRS theorists have themselves tended to forgo the analytic/synthetic distinction and regard *all* of an expression's inferential relations as constitutive. This "meaning holism" has the disturbing consequence that it would be virtually impossible for two people ever to mean exactly the same thing, indeed, for even one person to mean the same thing over any change of belief – rendering memory impossible! Fodor thinks that this renders any serious psychological generalizations impossible as well, and so is at pains to block the many arguments for it (1992a), and for any IRS, which, he thinks, inevitably invites it.

In Fodor's view, the original sin endemic to IRS theories consists in conflating *semantics* (or a theory of the content of concepts) with *epistemology* (or a theory of how we apply concepts). This conflation not only burdens semantics with the notorious problems of a verifiability theory of meaning that lurks in most of the above IRS proposals, but also presents substantial problems for accounting for the aforementioned productivity and systematicity of thought. Fodor argues that these latter phenomena require a compositional semantics (i.e. one in which the meaning of a complex expression is a function of the meaning of its parts), and epistemological capacities are not in general compositional: one could know a lot about pets and a lot about fish without knowing much at all about pet fish, for example, what one typically looks like (1990b, 1998b: chs 4–5). Unconfounding epistemology and semantics, Fodor instead forgoes any "molecular" account of meaning that depends upon relations among symbols, and instead embraces an "atomistic" theory that requires only that a symbol stand in a specific relation to the external world.

Information theories

Fodor takes as his point of departure the "information" theoretic semantics developed by Fred Dretske, which treats semantic meaning as a species of "natural" meaning (whereby dark clouds "mean" rain, or smoke fire). This idea often appears in psychological discussions under the guise of *discrimination abilities*: for example, something is a "shape receptor" if and only if it reliably discriminates shapes. The idea is naturally spelt out in terms of certain *counterfactual* dispositional properties to co-vary with specific phenomena in the world.

So stated, information semantics is open to several immediate objections.

1 pan-semanticism: something needs to be said about what's special about semantic or psychological meaning, since everything is causally related (and so "carries information") about *something*;
2 transitivity: "information" is transitive, but meaning isn't: if "smoke" co-varies with smoke, and smoke, itself, with fire, then "smoke" co-varies with fire; but "smoke" doesn't mean fire (1990b: 93);
3 robustness: *most* tokenings of sentences are produced in the *absence* of the conditions that they nevertheless mean. "That's a horse" can be uttered on a dark night in the presence of a cow, or just idly in the presence of anything. Fodor (1987b) calls these latter usages "wild," and the property whereby tokens of symbols mean things that aren't on occasion their actual cause, "robustness."
4 In accounting for robustness, a semantic theory needs to say what distinguishes the "wild" from the meaning-constitutive causes, a problem made vivid by the "disjunction" problem: what makes it true that some symbol "F" means HORSE and not HORSE OR COW ON A DARK NIGHT, or HORSE OR COW ON A DARK NIGHT OR W^2 OR W^3 OR . . . (where each W^i is one of the purportedly "wild" causes) (1987b).

A fifth problem could be raised regarding the contents of logical and mathematical symbols, which do not obviously enter into causal relations with any worldly phenomenon. Fodor sets aside this problem for the nonce, although suspecting that they are the only symbols for which an IRS is plausible.

Teleological views

A natural suggestion regarding the meaning-constitutive conditions is that they are in some sense "optimal" conditions that obtain when nothing (e.g. poor vision, limited spatiotemporal access) is "interfering" with belief fixation, and it is functioning as it was "designed to." Fodor (1987b) calls such theories "teleological" and he himself proposed a version of one in the widely circulated paper called "Psychosemantics" (1990a) (to be distinguished from the book (1987b), of the same title in which he *rejects* any such theory!). The attraction of such a theory lies in its capturing the idea that two individuals *meaning* the same thing by some symbol consists in their agreeing about what it would apply to, *were they to agree about everything else*. Their disagreements are to be explained as due to their differing epistemic positions and reasoning capacities.

Although Fodor nowhere suggests such theories are false, he does think they are subject to a number of difficulties, the chief one consisting of the circularity that seems

unavoidable in specifying the optimal conditions: it would appear that those conditions cannot be specified without employing the very intentional idiom the theory is supposed to explain (1987b: 104–6).

In order to avoid this and other problems Fodor (1987b, 1990b) went on to propose his "asymmetric dependency" theory. Although it makes no explicit appeal to ideal epistemic conditions, much of its motivation can be appreciated by thinking of the ideal co-variational theory in the background.

Asymmetric dependency

According to the ideal co-variational theory, tokens of an expression may be "wild," that is, produced by a property it doesn't express. Now, one way to understand the asymmetric dependency theory is first to notice that, plausibly, *all such wild cases depend upon the ideal case, but not vice versa*: the wild tokenings depend upon the ideal ones, but the ideal ones don't depend upon the wild ones. Getting things wrong depends upon getting things right in a way that getting things right doesn't depend upon getting things wrong. Thus, the property HORSE causes "cow" because some horses, for example, those at the far end of the meadow, look like cows and, under ideal conditions, COW causes "cow."

So formulated, of course, the account still mentions ideal conditions, and these Fodor has conceded cannot be specified without circularity. His further interesting suggestion is that mention of the ideal conditions here is entirely inessential: *the structure of asymmetric causal dependency alone, abstracted from any specific conditions or causal chains, will do all the required work* (1990b: 99, 1998a: 156ff).

To simplify the discussion, we can define a predicate, "*x* is locked onto *y*," to capture this asymmetric causal structure:

A symbol "S" is locked onto property F just in case:

1 there's a (*ceteris paribus*) law that F causes tokenings of "S";
2 tokenings of "S" are robust: i.e. are sometimes caused by a property G other than F;
3 when Gs (other than Fs) cause tokenings of "S," then their doing so asymmetrically depends on (1) i.e. on the law that F causes "S"s,

where X's causing Ys "asymmetrically depends" on a law, L, if and only if X's causing Y wouldn't hold but for L's holding, but not vice versa: L could hold without X's causing Y. Thus, smoking's causing cancer, depending upon many laws, asymmetrically depends upon Newton's, since Newton's doesn't depend upon smoking's causing cancer. Fodor's proposal about content then is:

(M) if "S" is locked onto F, then "S" expresses F.

Thus, a predicate "C" expresses COW if (a) it were a law that the property COW causes "C" tokenings, and (b) other causal relations between properties (e.g. HORSE, MILK, etc.) and "C" tokenings asymmetrically depend upon this law.

Note that (M) supplies only a *sufficient*, not a necessary physicalistic condition for predicate expression. Fodor believes this is all that he is required to do, given his merely "supervenient" physicalism mentioned on p. 453. Fodor argues that, if there are no

counterexamples to (M), then he has done all that he needs to do to show that, contrary to dualism, certain physical arrangements are sufficient for intentionality.

Note also that Fodor avails himself of the convenient largesse (some might regard it as a profligacy) of properties in the world. For him, as for many philosophers, there's virtually a property for every primitive predicate, *whether or not the property happens to be instantiated*. Thus, there are properties of being a unicorn and being phlogiston, despite the lack of any instantiations of them in the actual world. And so the concepts UNICORN and PHLOGISTON are distinguishable in this way by the respective lockings.

Fodor (1987b, 1990b, 1991b) defends (M) with considerable ingenuity. Whether or not this atomistic account of meaning can succeed, it is crucial to understanding Fodor's later work where it is simply taken for granted. (Fodor has written almost nothing further on (M) since 1991.)

Solipsism and narrow content

Despite the attractions of an "externalist" theory like (M), it is hard to resist the idea that there is *something* semantic purely "in the head." There are two standard ways of pressing this point: so-called "Frege" cases, and "Twin" cases.

Frege cases

There seem to be plenty of expressions with the same worldly reference that nevertheless have patently different meanings. Frege's example was "the morning star" and "the evening star," but these are distinguishable in different possible worlds, and so involve different properties. But what about predicates that are necessarily co-extensive, not only in this but all possible worlds, like "equilateral" versus "equiangular" triangle? Here Fodor avails himself of the resources of the LOT: "equi-angu-lar" and "equi-lateral" are syntactically complex, and so thoughts involving them can be distinguished thereby (1998a: 15–21, 163–5). In his terminology, they are different concepts with the same content.

Can all cases be handled in these ways? Are all differences in thought either differences in the denoted properties or structural differences in the way the properties are represented? Proper names present one kind of problem; terms for kinds ("lawyer," "attorney") another; necessarily co-instantiated terms, such as Quine's notorious "rabbit/undetached-rabbit-parts" example, still another. Fodor claims that, so long as coreferential names and simple kind terms are treated as tokens of different types *internally* by any agent, *inter*agent comparisons are merely pragmatically different (1994: 109–12), a view that marks a change from his earlier view and which he shares with many "direct reference" theorists such as David Kaplan. He also provides a detailed response to the Quinian challenge, exploiting a distinction in the logical role the different co-instantiated thoughts play (1994: 55–79).

Twin cases

Twin cases are the converse of Frege cases: instead of two expressions with the same reference but different senses, here we have expressions with the same sense but different references. Hilary Putnam invited us to imagine there was a faraway planet,

"Twin Earth," exactly like Earth in every way (including history) except for having in place of H_2O a superficially similar, but atomically different chemical XYZ (see PUTNAM). Oscar on Earth thinks about water, i.e. H_2O, where his twin on Twin Earth doesn't think about water at all, but about twater, i.e. XYZ. The question now is whether psychology should care about distinguishing Oscar from his twin: after all, aren't their internal mental lives indistinguishable?

Fodor's views about twin cases have changed over the years. In a much discussed paper (1980a) he argued that psychology couldn't wait upon a full theory of all an agent's environment, telling us which was water and which twater. And so it had better adopt what Putnam called a policy of "methodological solipsism," theorizing only about what goes on inside an agent's head. He takes this to comport well with what he calls a "formality condition" that follows from CRTT: mental states have their efficacy as a consequence of the formal character of their tokens.

However, he also recognizes that intentional properties of mental representations are essential to their role in psychological explanation, and so, if psychology is solipsistic, there must be some solipsistic, or "narrow" kind of content that supervenes on the internal states of a thinker.

For a while, Fodor settled on the following conception (originally proposed by Stephen White, developing the seminal work of Kaplan): *the narrow content of a Mentalese expression is a function (in the set theoretic sense) that maps a person's life context onto a broad content*. For example, the narrow content of Oscar and Twin Oscar's "water" is the function that maps Oscar's context onto H_2O and his twin's context onto XYZ. When Oscar utters "Water is wet," he thereby expresses the content "H_2O is wet," while when Twin Oscar utters it he expresses the content "XYZ is wet." Two symbols have the same narrow content just in case they serve to compute the same such function: it is this that is shared by Oscar and his twin. Whether this function can actually be specified by psychology is, however, not altogether clear: how does one continue to specify it beyond Earth and the fanciful case of Twin Earth?

Only wide content after all

More recently Fodor has moved away from any reliance on narrow content at all. He argues that both Frege cases and Twin cases don't have to be taken seriously by psychology: they are violations of the *ceteris paribus* conditions under which serious psychological laws are satisfied (1994: ch. 2). With respect to the Twin cases, he claims that it's important to remember what he earlier forgot, that, although they are *conceptually* possible, they are not *nomologically* so, and "empirical theories are responsible only to generalizations that hold in nomologically possible worlds" (1994: 29).

With respect to the Frege cases, his position is more complex. Citing the case of belief, he argues for what he calls the "Principle of Informational Equilibrium" (PIE),

> Agents are normally in *epistemic equilibrium* in respect of the facts on which they act. Having *all* the information – having all the information that God has – would not normally cause an agent to act otherwise than as he does. (1994: 42)

He claims that, since the success of our action is no accident and tends to depend upon the truth of our beliefs, "no belief/desire psychology can fail to accept PIE" (1994: 42).

Consequently, Frege cases, in which an action depends upon being *ignorant* of an identity statement are, from the point of view of psychology, "aberrations." They do occur; and, following the earlier discussion, they can be described by invoking syntactically different LOT expressions as the different "modes of presentation" Frege thought were needed. But, *pace* Frege and his many followers, no "third realm" of "senses," or narrow contents, is needed. Whether a similar argument can be made for attitudes other than belief is a question that is unfortunately not addressed.

Nativism

In the same book in which Fodor presented the CRTT hypothesis he also defended a highly controversial thesis about *the innateness of all concepts*. That thesis originally stirred more controversy than did CRTT itself, much of which was addressed in publication (1981c); but it is independent of CRTT, depending upon further claims about the nature of concepts, definitions, and learning, issues which, moreover, are not likely to be settled by commonsense thought on the matter (1998a: 28).

In (1975a: ch. 2) Fodor argued that, since standard models of learning involved hypothesis confirmation – one learns that "cat" in English means *cat* by confirming hypotheses about English usage – these models are committed to the constituent concepts of these hypotheses being innate (1975: ch. 2): after all, one can't form hypotheses such as " 'red' means *red*," without using and therefore already possessing those very concepts!

Later he deepens the discussion by considering the traditional empiricist suggestion that one acquires concepts by constructing complex ideas out of sensory primitives that are variously associated in experience (1981b). The persistent failure over the centuries to set out these constructions, which is closely related to the failures of any IRS (see section "Inferential role theories"), suggests that they are probably not to be had. As Plato observed early on (and Chomsky more recently), most of our cognitive capacities seem to transcend our specific sensory histories.

However, Fodor (1998a: chs 6–7) has recently raised what he regards as a serious objection to his earlier views that all concepts are innate, what he calls the "DOORKNOB/'doorknob' " problem:

> Why is it so often experiences of doorknobs, and so rarely experience with whipped cream or giraffes, that leads one to lock onto *doorknobhood*? . . . assuming that primitive concepts are triggered, or that they're "caught," won't account for their content relation to their causes; apparently only induction will. But primitive concepts can't be induced; to suppose they are is circular. (1998a: 127–32)

That is, as he argued in 1975a: ch. 2, you can't perform inductions to concepts that you can't already represent.

His solution to this problem is "ontological": the properties that correspond to our primitive concepts are just the properties to which we generalize, *from phenomenally specified stereotypes*. For example, it is constitutive of *being a doorknob* that it is the property onto which people lock as a result of exposure to stereotypical doorknobs. Fodor relies here on what he regards as the psychological evidence that stereotypical instances

of primitive concepts can be specified *independently of those concepts*, e.g. by enumerating the shapes, colors, functions that typical instances share (1998a: 137–45).

Fodor thinks we now have an acceptable account of the non-arbitrary relation between the acquisition of many of our concepts and the experience of typical instances of them. Moreover, it has the interesting consequence that someone *not exposed to typical doorknobs* might well not have the concept "doorknob," since, without that experience, there might well be no such locking. He speculates that this is the case for most commonsense concepts. So, he concludes, "maybe there aren't any innate *ideas* after all" (1998a: 143). All there are are innate dispositions to lock onto properties when exposed to their stereotypical instances.

There are a number of problems raised by this view. It might appear to undermine the reality of the referents of our commonsense concepts, making them "dependent upon us." But this is an illusion. A doorknob may be identified by its tendency to have a certain effect on us; but it can exist even if, if we didn't exist, it didn't have that effect. But, in any case, minds are in fact a part of the world, and doorknobs do in fact have their effects upon them (1998a: 148–9).

A more serious problem is raised by what seem to be genuine "natural kind" concepts, whose reference patently does not involve any relation to what people might do: being genuine *water* is a matter of being H_2O, whether or not people would generalize to it on the basis of stereotypical samples. These concepts, Fodor claims, are latecomers in our cognitive development, dependent upon the social institution of science, the development of sophisticated theories and the consequent deference to experts (1998a: 150–62). His hope is that the peculiar semantic effects of these late developments can be entirely spelt out in terms of details of the counterfactuals involved in his asymmetric dependence theory (see "Teleological views").

Contra selectionism

Fodor's interest in nativist hypotheses might lead one to think that he believes that our cognitive capacities are the result of natural selection. Nothing could be further from the truth. Although he doesn't doubt for a moment that the human mind/brain evolved, he sees no reason whatever to think that its considerable cognitive capacities were specifically *selected*. In part his opposition to selectionism is like his opposition to empiricist theories of learning: he sees no reason to think that these capacities reflect some sort of regularities in our histories. Many of them – such as our grammatical or mathematical abilities – seem to far exceed anything that either our upbringing or our evolutionary history could plausibly have supported.

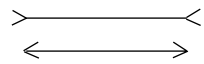
Additionally, he thinks that selectionist stories about the evolution of cognitive capacities, beside being flagrantly speculative, seriously underestimate the complexity of the relation between mind and brain: "make an ape's brain just a little bit bigger (or denser, or more folded . . .) and it's anyone's guess what happens to the creature's behavioral repertoire" (1998b: 209). It's as likely as not that some small change in our ancestors' brains made them *tremendously* smarter, like the modest change required to transform a finite state machine into a Turing machine. Doubtless this provided us with *some* selectional advantages. But that is no reason to suppose that anything like the majority of mental abilities we display – from acquisition of grammars to the grasp

of higher mathematics, physics, or folk psychology – were themselves individually selected.

Modularity and the limits of CRTT

In discussing IRS theories in the section "Inferential role theories," we noted their tendency to become "holistic," a tendency Fodor sees as inimical to the interest of serious psychology. This tendency has also been prevalent in much contemporary "New Look" theories of perception: the work of Jerome Bruner in psychology, Thomas Kuhn in the history of science (see KUHN), and Nelson Goodman in philosophy (see GOODMAN), emphasizes how much people's theoretical expectations can color their perceptions, to the point that they and others insist that we ought to abandon "the myth of the given" (see also SELLARS). Fodor deplores this holistic tendency here as well, and in this case marshals interesting psychological evidence against it.

Fodor first of all calls attention to the surprisingly little noticed fact that the very perceptual illusions that New Look theorists often invoked to make their point actually tell against it: for *many of these illusions do not disappear when we know better*. No matter how sure we are that the Muller–Lyer illusion lines are equal, the upper still looks longer



than the lower. In a phrase Fodor takes from the psychologist, Zenon Pylyshyn, perception seems to be "cognitively impenetrable."

Fodor cites facts like these, and considerable data about language comprehension, to argue that there are a number of dedicated mental "modules" that are "informationally encapsulated" from the "central" system whereby we reason generally and fix our beliefs. These include the standard sensory systems, certain levels of language processing, and perhaps other dedicated systems such as face and musical perception. Among other things, such systems are, furthermore: *extremely rapid* (on the order of a quarter of a second, 1983: 61–4); *shallow* (their outputs are limited to "basic perceptual categories," such as chair or dog, 1983: 86–99); associated with a *characteristic development* (vision and language seem to develop in specific ways that are independent of other mental capacities, 1983: 100–1); and they are *domain specific* (confined to, for example, processing of light, faces, or grammar, 1983: 47–52). In this last respect, Fodor's conjecture overlaps, of course, with Chomsky's postulation of an innate grammatical competence, but, further, involves the modularity of not only the domain, but of the standard processing of information from that domain (1989, 1998b: ch. 4).

All of these properties contrast with properties of the central system, which is *voluntary* (you can choose what to think about), *slow* (you can potter for months), sometimes *deep* (you can think about non-perceptual categories), non-localized ("there is, to put it crudely, no known brain center for *modus ponens*," 1983: 98), and *domain independent* (you can think about almost anything).

Philosophically, what is perhaps most intriguing about the postulation of sensory modules is the way in which it provides a new basis for the controversial observation/theoretic distinction, although a basis that may only partially overlap the traditional introspective one (1984b, 1988b).

It might be thought that, given their encapsulation, these modules are not really *cognitive*, and so don't involve all the issues about computation and representation that are the focus of most of Fodor's work. But precisely the point of postulating modules as opposed to standard transducers is that they do seem to involve computation. Indeed, Fodor regards them as "compiled transducers. . . . 'compiled' to indicate that they have an internal computational structure, and 'transducer' . . . to indicate . . . information encapsulation" (1983: 41).

Indeed, in his recent work (2000), Fodor argues that modules may be the only appropriate domain for a CRTT. In being *un*encapsulated, the *central* belief fixation system exhibits a number of properties that present serious prima-facie difficulties for any standard computational treatment. Relying on what he regards as Quine's astute views about theory confirmation, Fodor claims that central systems are:

1 "*Quinian*," i.e. computed over the totality of a belief set, as when we settle on a theory that is, for example, simplest and most conservative overall;
2 *isotropic* (every belief is potentially relevant to the confirmation of every other, as when a pattern of light on a piece of paper confirms a theory about the age of the universe) (1983: 105ff).

Fodor (1999) argues that these features render belief fixation holistic and context-dependent, in a fashion that is not clearly amenable to the Turing computability invoked by CRTT. This latter depends upon exploiting a representation's local syntactic features: an argument's deductive validity can be checked by looking at its local spelling. However, abductive cogency seems to be ascertainable only by looking at a claim's relation to other, indefinitely remote representations, and its effect on the belief system as a whole. Fodor sees this as the problem underlying the so-called "frame problem" encountered in artificial intelligence (1987a), and is consequently pessimistic about the prospects of it being ultimately solved by CRTT. Although CRTT is necessary for an adequate theory of mind, it seems to be far from sufficient.

Bibliography of works by Fodor

For a full bibliography of Fodor's work up until 1991, see B. Loewer and G. Rey (1991) *Meaning in Mind: Fodor and his Critics* (Oxford: Blackwell Publishers), which also contains critical essays by a number of prominent philosophers, and an introduction on portions of which the present entry relied.

1963 (with Katz, Jerrold): "The Structure of a Semantic Theory," *Language* 39, pp. 170–210.
1968a: *Psychological Explanation*, New York: Random House.
1968b: "The Appeal to Tacit Knowledge in Psychological Explanation," *Journal of Philosophy* 65, pp. 627–40.
1970: "Three Reasons for not Deriving 'Kill' from 'Cause to Die'," *Linguistic Inquiry* 1, pp. 429–38.
1972 (with Block, N.): "What Psychological States are Not," *Philosophical Review* 81, pp. 159–81.
1974 (with Bever, T. and Garrett, M.): *The Psychology of Language*, New York: McGraw Hill.
1975a: *The Language of Thought*, New York: Thomas Y. Crowell.
1975b (with Fodor, J. D. and Garrett, M.): "The Psychological Unreality of Semantic Representations," *Linguistic Inquiry* 6, pp. 515–31.

1978: "Tom Swift and his Procedural Grandmother," *Cognition* 6, pp. 229–47.
1980a: "Methodological Solipsism Considered as a Research Strategy in Cognitive Science," *Behavioral and Brain Sciences* 3, pp. 63–109 (with replies to commentators).
1980b (with Garrett, M., Walker, E. and Parkes, C.): "Against Definitions," *Cognition* 8, pp. 263–367.
1981a (with Pylyshyn, Z.): "How Direct is Visual Perception? Some Reflections on Gibson's 'Ecological Approach'," *Cognition* 9, pp. 139–96.
1981b: "The Present Status of the Innateness Controversy," in Fodor 1981c, pp. 257–316.
1981c: *Representations: Essays on the Foundations of Cognitive Science*, Cambridge, MA: MIT Press.
1983: *The Modularity of Mind: An Essay on Faculty Psychology*, Cambridge, MA: MIT Press.
1984a: "Semantics, Wisconsin Style," *Synthese* 59, pp. 231–50.
1984b: "Observation Reconsidered," *Philosophy of Science* 51, pp. 23–43.
1986: "Why Paramecia Don't Have Mental Representations," in *Midwest Studies in Philosophy*, vol. X, ed. P. French, T. Uehling, Jr., and H. Wettstein, University of Minnesota Press.
1987a: "Frames, Fridgeons, Sleeping Dogs and the Music of the Spheres," in Z. Pylyshyn (ed.) *The Robot's Dilemma; The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex.
1987b: *Psychosemantics: The Problem of Meaning in Philosophy of Mind*, Cambridge, MA: MIT Press.
1988a (with Pylyshyn, Z.): "Connectionism and Cognitive Architecture," *Cognition* 28/1–2, pp. 3–71.
1988b: "A Reply to Churchland's 'Perceptual Plasticity and Theoretical Neutrality'," *Philosophy of Science* 55, pp. 188–98.
1989: "Why Should the Mind be Modular?," in A. George (ed.) *Reflections on Chomsky*, Oxford: Blackwell Publishers.
1990a: "Psychosemantics, or Where do Truth Conditions Come From," in W. Lycan (ed.) *Mind and Cognition*, Oxford: Blackwell Publishers.
1990b: *A Theory of Content and other Essays*, Cambridge, MA: MIT Press.
1991a (with McLaughlin, B.): "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work," *Cognition* 35/2, pp. 185–204.
1991b: "Replies," in *Meaning in Mind: Fodor and his Critics*, ed. B. Loewer and G. Rey, Oxford: Blackwell Publishers, pp. 255–319.
1991c: "You Can Fool All of the People Some of the Time, Everything Else Being Equal: Hedged Laws in Psychological Explanations," *Mind* 100/1: pp. 19–34.
1992a (with LePore, E.): *Holism: A Shopper's Guide*, Oxford: Blackwell Publishers.
1992b: "Substitution Arguments and the Individuation of Beliefs," in G. Boolos (ed.) *Essays for Hilary Putnam*, Oxford: Blackwell Publishers.
1994: *The Elm and the Expert*, Cambridge, MA: MIT Press.
1998a: *Concepts: Where Cognitive Psychology Went Wrong*, Oxford: Oxford University Press.
1998b: *In Critical Condition*, Cambridge, MA: MIT Press.
2000: *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, Cambridge, MA: MIT Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

38

Saul Kripke (1940–)

DAVID SOSA

Life

Kripke once said, "People used to talk about concepts more, and now they talk about words more. . . . Sometimes I think it's better to talk about concepts." In fact, Kripke himself has said important things, and developed and deployed significant conceptual resources, about both words and concepts.

Saul Aaron Kripke was born in Bay Shore, New York. His mother Dorothy was a teacher and father Myer a rabbi. The family soon moved to Omaha, Nebraska where Kripke spent most of his childhood. He was a child prodigy, learning Hebrew on his own at the age of 6 and reading all of Shakespeare in the fourth grade. But it was in mathematics that he exhibited the greatest precocity: he derived results in algebra – intuitively, without the benefit of algebraic notation – in fourth grade and taught himself geometry and calculus by the end of elementary school. By the time he was in high school, Kripke's work in mathematical logic was so advanced that he presented some of it at a professional mathematics conference. Around the time he published his first article, "A Completeness Theorem in Modal Logic," Kripke was on his way to Harvard, from which he graduated with a bachelor's degree in mathematics in 1962. But during his years at Harvard, Kripke's interests already began to shift to philosophy.

In 1963 Kripke was appointed to the Harvard Society of Fellows and later to positions as lecturer at Princeton University (1965, 1966) and back at Harvard (1966–8). Finally, he was appointed Associate Professor at Rockefeller University in 1968 and promoted to Professor in 1972. But the outstanding philosophy department at Rockefeller was disbanded (by the University's President, Frederick Seitz) in the mid-1970s and Kripke was appointed McCosh Professor of Philosophy at Princeton in 1977, the position from which he retired in 1999.

Modal logic

Early in his career, Kripke made essential and seminal contributions to modal logic. Modal logic is, in effect, the logic of necessity and possibility and its history can be traced to at least Aristotle. In the first half of the twentieth century, C. I. Lewis, C. H. Langford, and then Carnap revived and developed modal logic. Lewis criticized the logical system

Russell and Whitehead had proposed in *Principia Mathematica* (which could not distinguish what is simply false from what is *necessarily* false – that is, what is *impossible*). With Langford, Lewis described five different axiom systems that could represent a new concept of logical entailment: *strict* implication. Unlike the notion of implication formalized in *Principia*, *p* does not get to strictly imply *q* simply in virtue of being false: it has to be impossible for *p* to be true and *q* false. Carnap later characterized the sort of logical necessity involved in strict implication in terms of truth in all "state descriptions." But Kripke, with his "Kripke models," made this idea of necessity precise, refined it, and generalized it.

Kripke models involve a set of "possible" worlds and, for each world, an assignment of truth-values to simple ("atomic") sentences. As developed by Kripke, this system enables us to characterize the notion of logical necessity that Carnap discussed (see CARNAP): necessary truths are those that are true at *all* possible worlds in *every* model. By including in addition an "accessibility relation" (meant to select the worlds that are possible *relative* to any given possible world), Kripke was able flexibly and systematically to characterize many other modal logics that are weaker than that suggested by Carnap's discussion. Indeed, much of the later progress of modal logic has depended on the idea of Kripke models, as well as on the notion of "Kripke frames," which are just like Kripke models (specifying a set of possible worlds and an accessibility relation) but without the evaluation of atomic sentences.

Meaning

After this important work in modal logic, Kripke turned his attention to the philosophy of language, revolutionizing that field with a series of publications in the period between 1971 and 1982. In "Identity and Necessity" and the early, article version of "Naming and Necessity," Kripke begins to develop the exciting ideas and arguments that get their fullest treatment in the book version of *Naming and Necessity* in 1980. These works challenge long-held assumptions about meaning while rehabilitating others, offer a new paradigm (or "picture," to use Kripke's term) of reference and meaning, and propose, on the basis of the developing theory of meaning, provocative theses in metaphysics, epistemology, and philosophy of mind.

Fundamentally, Kripke argues that a traditional view of meaning is mistaken. In the tradition Kripke sees as beginning with Frege and Russell (see FREGE and RUSSELL), names, for example, refer to what they do in virtue of being associated with some descriptive content. The referent of the name is what *satisfies* the descriptive content associated with it. With a name such as, say, "Aristotle," one might think the descriptive content would include *taught Alexander* or *was the student of Plato*, and so on. Kripke presents, in compelling form, a battery of arguments against any such view. These arguments can profitably be seen as coming in three varieties: (1) modal, (2) semantic, and (3) epistemic.

The modal argument begins with an observation for which Kripke is now celebrated: names are *rigid designators*. A rigid designator is a word that designates the same object with respect to any possible situation. So, for example, we may say that if he had been chosen to lead the Academy, Aristotle would never have gone on to teach Alexander. When we make that statement, it's a claim about a situation (or what can also be called

a "possible world") that's different from our own; in our world, Aristotle was *not* chosen to lead the Academy after Plato. But even though we're talking about a different situation, we're talking about *Aristotle* in that situation. So the name "Aristotle" maintains its reference to Aristotle, even with respect to possible situations in which Aristotle was chosen to head up the academy, did not go on to teach Alexander, or in which his life varied in any of the ways it might have.

But notice that since names are rigid designators, we can make true claims about what *might* have happened that would appear to be ruled out by the description theory Kripke opposes. Consider any description we might think is part of the descriptive content of the name "Aristotle": say, *was born in Stagira*. Aristotle, of course, might have been born elsewhere, if his parents had moved before he was born, for example. On the other hand, no one can both be born and not be born in Stagira. So while the sentence, "Aristotle was not born in Stagira," seems to express something that's possible, any sentence like "The . . . who was born in Stagira was not born in Stagira," seems to express something impossible. But if part of what "Aristotle" *means* is *was born in Stagira*, then it's hard to see why these two sentences should differ in this way. Why is what's expressed by one sentence possible and what's expressed by the other impossible, when they have, relevantly, the same meaning? This is an example of Kripke's modal argument.

There are several other ways of putting the point of the modal argument. But they can be seen as reducing to a general pattern: names are rigid designators, descriptions are not; therefore descriptions cannot give the meaning of names (in the way proposed by the traditional view of Frege and Russell). Names have a different modal *profile* from descriptions.

Even if Kripke had given none other, many would find the modal argument sufficiently devastating to refute the traditional view of names at which it's directed. But an important part of the significance of Kripke's work on meaning is that he presents, as noted above, a battery of arguments, each of which is a further, independent point against the traditional view he challenges.

Kripke's epistemic argument has a structure similar to that of his modal argument. If *was a student of Plato's* is literally part of the meaning of the name "Aristotle," then we should expect the sentence "Aristotle was a student of Plato's" to express a trivial a priori truth that could be known without any historical or empirical investigations. But you might be a competent user of the name "Aristotle" without knowing that Aristotle was a student of Plato's. Perhaps all you know is that Aristotle was some great philosopher. The description theory predicts that certain sentences should be a priori when in reality they are not.

And Kripke's semantic arguments suggest that the referent of a name is not whatever satisfies the descriptions that might be associated with it. He is aided here by compelling examples. In one, Kripke asks us to imagine a circumstance in which Kurt Gödel did not discover the incompleteness of arithmetic (as, in fact, he did), but rather stole that result from someone named "Schmidt." Now, it's plausible that something like "discovered the incompleteness of arithmetic" is associated with the name "Gödel." But notice that in this case, that would yield Schmidt as the referent of the name "Gödel." Kripke uses this as an argument against the description theory. Surely even with respect to a situation in which Schmidt is the discoverer of the incompleteness of arithmetic,

"Gödel" refers to Gödel and not to Schmidt. But that means the name "Gödel" is not tied to its referent by means of the satisfaction of the description *discovered the incompleteness of arithmetic*. If the meaning of "Gödel" were the descriptive content associated with it, then the name would refer to the wrong person – it would have the wrong semantics. Another example: *a famous physicist* picks out Gell-Mann as much as it does Feynman. Still, even if that's the only descriptive content associated with the name "Feynman," the name refers to Feynman and not to Gell-Mann.

Acknowledging a debt to J. S. Mill, Kripke holds that names are *denotative* but *non-connotative*. The meaning of a name is exhausted by its referent. Rather than having any descriptive content as its meaning, a descriptive content that would then determine a referent, Kripke suggests that the meaning of the name just *is* the referent itself. This claim is now considered constitutive of a position known as "Millianism" in philosophy of language.

This leaves open the question of why a name has the referent it does. In place of the description theory he associates with Frege and Russell, Kripke offers an alternative "picture" of the naming relation. In the causal account he suggests (sometimes called the "historical chain" account), a name has the meaning it does – that is, it *refers* as it does – in virtue of a chain of causal relations between uses of the name and the referent. Kripke explicitly admits not having anything like a "theory": but he proposes causation as the fundamental mechanism by which reference is fixed (though these causal relations do not themselves constitute the meaning; the meaning, recall, just *is* the referent).

It's an interesting fact that although he attacks a descriptive theory of naming associated with Russell, in other work Kripke ingeniously defends Russell's theory of descriptions themselves. According to Kripke, Russell was wrong to view names on the model of descriptions; but his account of descriptions themselves was unobjectionable. Russell's theory of descriptions (in "On Denoting") concerned the meaning of expressions such as "the President" or "The even prime number," or even "Plato's most famous student." In 1966, the philosopher Keith Donnellan issued a challenge with an example in which a sentence containing a definite description seemed to have a meaning that was inconsistent with what would be predicted by Russell's theory. Drawing on a distinction between language use and language meaning, and distinguishing between *speaker reference* and *semantic reference*, Kripke answers Donnellan's challenge and defends Russell's theory of descriptions.

One serious problem for the sort of theory Kripke's arguments support (though, again, Kripke himself never explicitly adopts any particular "theory") concerns belief and belief ascription. If names are merely denotative and are non-connotative, then, since the meaning of a name is exhausted by its referent, any two names with the same referent have the same meaning. But given just a few other plausible assumptions, this entails that there should be no difference in meaning (and thus no difference in truth-value) between sentences like "Lois Lane believes Clark Kent can fly" and "Lois Lane believes Superman can fly." But (among other problems) it seems that what Lois really believes is that Clark Kent cannot fly.

In his "A Puzzle About Belief" (1979), Kripke argues that this unwelcome result is not due to any features specific to the position in question: our practices of belief ascription themselves, independent of any specific assumptions about the meaning of names,

will yield the same unwelcome results. He uses the now-infamous (in philosophy of mind and language!) example of Pierre, a normal monolingual Frenchman, who hears of that famous distant city, London (which Pierre of course calls "Londres"). On the basis of what he has heard of London, he is inclined to say, in French, "Londres est jolie." Taking him at his word, and translating, we can conclude that he believes that London is pretty. Later, Pierre leaves France and moves to an unattractive part of London. He learns English by the "direct method," without using any translation between English and French. Pierre is unimpressed with his surroundings and is inclined to assent to the English sentence "London is not pretty." Again, taking him at his word, we can conclude that he believes that London is not pretty. But now he seems to be in much the same position as Lois above.

What's important, for Kripke's purposes, is that we seem to have put Pierre into that position without explicitly appealing to a "Millian" (names are merely denotative) position. That suggests Millianism is not a distinctively problematic position. The sort of puzzle that's put forward against Millianism is really a problem for everyone, Kripke argues. Thus he defends Millianism from its main challenge.

Necessity, a priority, the mind–body problem, and essentialism

Kripke's revolution in philosophy of language would have been more than enough to secure his importance. But Kripke went on to transform his theses about meaning into interesting positions in metaphysics, epistemology, and philosophy of mind. Perhaps the most significant element of his meaning theory, for these purposes, is his distinction between what "fixes the reference" of a term (which for names, he suggests, is typically fundamentally a causal relation) and the actual meaning of that term (which, in the case of names, consists of the referent itself).

Since Immanuel Kant in the late 1700s, philosophers had traditionally seen two sorts of phenomena as intimately related. A proposition was taken to be *necessary* if it cannot possibly fail to be true, and counted as a priori if, roughly, it can be known without the benefit of empirical investigation. It was natural to think that all necessary propositions are a priori and that, with a few special exceptions, those that are not necessary can be known only a posteriori. If a proposition is necessary, then one needn't see how the world is as a matter of fact in order to know that proposition. Its truth does not depend on the state of the world; empirical investigation thus seems beside the point. And, conversely, if a proposition is contingent, then how could it be known a priori? Since it's not true in *every* possible world, we would have to investigate the world around us to see whether it's true in ours. (One exception is Descartes's *Cogito* – I think, therefore I am – whose premise, and conclusion, each seem contingent and yet, in one sense, a priori). Shockingly, Kripke rejected both directions of this alleged intimate relation.

According to Kripke, necessity and a priority are not nearly as intimately related as had been thought. There are necessary truths that can be known only a posteriori and a priori truths that are contingent. And these aren't just exceptional, unusual cases, but systematic, standard occurrences. Consider an example Kripke uses, picking up on a comment of Wittgenstein's, to support his claim that what's a priori can be contin-

gent. We introduced the word "meter" and fixed its reference with respect to a certain standard: the standard meter bar in Paris. (The reference has since been re-fixed, but set that aside.) Now take the claim that the standard meter is one meter long. How can we know this? The idea of *measuring* the standard meter is ludicrous: our knowledge that the standard meter bar is one meter long is not the sort of thing that is to be checked empirically. The standard meter is precisely what fixes the reference of the term "meter." But is it a *necessary* truth that the standard meter is one meter long? Kripke reminds us that the standard meter bar might have been longer than it in fact is. Indeed, if just before we fixed the meaning of our word "meter" with reference to that bar, it had undergone some significant temperature change (that it did not, as a matter of fact, undergo), then the bar would have been longer (or shorter) than a meter. Of course, in that circumstance, we'd use the word "meter" for that new length. But it's still true that the meter bar in that circumstance wouldn't be a meter long: we'd just be using the word "meter" for a different length. We know a priori that the standard meter's a meter long; but it *might* not have been. There are possible circumstances in which the standard meter bar has a different length.

Conversely, necessity does not entail a priority. Gold has atomic number 79 and water is H_2O. According to Kripke, these are not things we could have known a priori. The chemical composition of water and the atomic number of gold were empirical scientific discoveries. We used some superficial identifying marks to fix the references of our terms "water" and "gold." Now, those marks don't *define* the words, they don't give their meanings. They served to pick out kinds which we then investigated empirically. But it is through empirical investigation that one discovers gold's atomic weight and water's chemical composition. Nevertheless, Kripke thinks the statements "gold has atomic number 79" and "water is H_2O" are necessary. There's no possible circumstance in which *gold* has any atomic number other than 79; and *water* couldn't be anything but H_2O. There may be circumstances in which *what we call* – in those circumstances – "gold" has a different weight, or in which what we call "water" has a different chemistry, but those are just worlds in which we use the terms for other stuff. (Of course, that's not to say we'd be making a *mistake* in calling that other stuff "gold" or "water": in those other circumstances, the words wouldn't have the same meaning they actually have.) According to Kripke, it's a matter of necessity that water be H_2O and that gold have atomic number 79. Having those chemical natures is what makes water and gold what they are. Science can discover essences.

But Kripke wasn't finished yet. Before closing his work on these matters, he takes on two other shibboleths: (1) at the time he wrote *Naming and Necessity*, a popular response to the mind–body problem – the traditional philosophical problem of the nature of mind and its relation to the physical body – was a kind of "identity theory." The idea was to view the problem as solved by contemporary science in much the same way that contemporary science had discovered the nature of, for example, heat. We can suppose that heat was originally identified as what produces a certain distinctive sensation. Through empirical investigation, we find that it is the kinetic motion of molecules that produces those sensations. So, roughly, heat *is* the motion of molecules. The then-popular identity theory wanted to view the relation of mind to body as akin to that between temperature and mean molecular kinetic energy. As we investigate the brain further, and discover which states are correlated with which

mental phenomena, we learn what these mental states *are*, just as we learned what temperature *is*.

Take pain. The mental state of pain appears to be correlated with the stimulation of what are called "C-fibers." Is that just what pain *is*? Have we solved the mind–body problem? Kripke points out that if we were to take the mental state of pain to just *be* the stimulation of C-fibers, then that would constitute the empirical discovery of a *necessity*, on the model of the discovery of the chemical constitution of water (remember: "water is H_2O" is necessary) or the nature of temperature. But there's a problem. In these cases of theoretical identification, of the scientific discovery of necessity, there is an explanatory note to be paid off: what explains the *illusion* of contingency? For it certainly seems that water might have turned out not to be H_2O. As we were performing the chemical investigations, at least, it seemed to be a contingent matter, possibly turning out one way, possibly another.

There is a standard way to make good on this explanatory debt: the identifying marks by which we fixed the reference of the relevant terms are, indeed, only contingently related to the essence of the kinds. So being the colorless, odorless liquid that falls from the sky as rain, etc. – that set of properties by which we identify water – is only contingently related to being water. Water might have existed without having those identifying marks. So although water *must* be H_2O, it can seem as though it need not have been, because H_2O need not be a colorless, odorless liquid that falls from the sky as rain, etc. Similarly with heat. Heat is necessarily mean molecular kinetic energy; but it's not a necessary truth that mean molecular kinetic energy produces the *sensation* of heat. That sensation is just a mark that we used to identify the phenomenon to be investigated.

Now comes Kripke's insight: in the case of pain, there's no analogous move! The marks by which we identify pain are *essential* to it; pain could not exist without being *felt* as pain. So if the stimulation of C-fibers could occur without being felt as pain, this would refute the mind–brain identity theory. Pain appears to be only contingently related to the stimulation of C-fibers. The identity theory must, according to Kripke, deny that appearance as mistaken. But it cannot explain its plausibility as it does in the analogous cases. For the mark by which we identify pain, its painful feeling, is *essential* to pain.

That leads us to the other shibboleth Kripke attacked: anti-essentialism. (2) in the 1960s and into the 1970s, influenced by Quine among others, many philosophers were opposed to essentialism – belief in modality *de re* – while accepting modality *de dicto* (see QUINE; cf. MARCUS). In other words, it was widely accepted that statements could be necessarily or possibly true or false (modality *de dicto*) but widely denied that it made sense to speak of a particular individual's necessarily or only contingently having a given property (modality *de re*). Kripke argues that a material object's material origin (the stuff from which it was made) is essential to it: *it* could not have been made from anything else. And he argues that one has one's parents *essentially*, so that one could not have had different parents. These are *de re* necessities; properties that individuals have necessarily. It's true that the method and force of his argumentation here, as elsewhere, is largely intuitive; but Kripke holds that although "some philosophers think that something's having intuitive content is very inconclusive evidence in favor of anything" he himself doesn't "know, in a way, what more conclusive evidence one can have about anything, ultimately speaking" (1980: 42).

Truth

In his groundbreaking "Outline of a Theory of Truth," Kripke makes a number of important advances in our theoretical understanding of truth. The paper quickly became a focus of all subsequent discussions. A main problem for our understanding of truth is presented by the so-called "liar paradox." Consider sentence (1):

(1) Sentence (1) is false.

Is sentence (1) true or is it false? Well, exploiting the attractive idea that a sentence is true just in case what it says holds, we might suppose that sentence (1) – that is, "Sentence (1) is false" – is true just in case sentence (1) is false. But now we have a problem: for we are saying that sentence (1) is true if and only if it is false. Indeed, whether sentence (1) is true or false, it follows that it's *both* true and false!

Tarski confronted this paradox, or in effect a metalinguistic version of it, and concluded that languages for which the paradox arises are "inconsistent": they are languages in which a sentence and its negation are jointly true. He suggested that such languages were inadequate for a theory of truth and proposed replacing them with more regimented languages, whose rules prevented the paradox. Tarski proposed a hierarchy of languages, none of which contains a "truth predicate" that applies to sentences of that very language (at that same level of the hierarchy) (see TARSKI, CHURCH, GÖDEL). A truth predicate (for a language *L*) is any predicate *T* which makes the following schema true for all instances (where one obtains an instance of the schema by replacing "S" with a sentence of *L*):

$\ulcorner T \ulcorner \mathrm{S} \urcorner \urcorner$ is true if and only if $\ulcorner \mathrm{S} \urcorner$ is true.

By prohibiting the application of a truth predicate to sentences in the same language, Tarski prevents the construction of the liar paradox; but he gives up the idea that there can be a satisfactory theory of truth for English (which apparently does have a truth predicate – namely, "is true" – that applies to sentences in the same, English, language).

Kripke shows how, if we allow "truth-value gaps" (i.e. if we allow sentences that are neither true nor false) we can make progress. But we do not simply *eliminate* the paradox by alleging that (1) is neither true nor false, because we can readily see that the paradox will rearise, in strengthened form, with the sentence:

(1′) Sentence (1′) is not true.

Perhaps (1′) is neither true nor false; but then it's not true. In which case, since what it says is that it's not true, it must be true. So the paradox rearises even if we allow truth-value gaps.

To make progress, Kripke introduces the notion of a "grounded" sentence with reference to the notion of a "fixed point" (which, very roughly, is an interpreted language whose interpretation assigns to the truth predicate all and only the true sentences of that language). This notion of groundedness is useful because, according to the constructive procedure by which it is understood, not every sentence of a language will be grounded. Some sentences, like (1) or (1′), may be ungrounded, not part of the extension of the candidate truth predicate, but also not part of the anti-extension either (where the anti-extension includes just those sentences to which, according to the

interpretation, the truth predicate does *not* apply). The point is that the interpretation may be *partial*, some sentences characterized as ones to which the candidate truth predicate applies, others characterized as sentences to which the predicate does not apply, and others simply left uncharacterized. Kripke suggests that an ungrounded sentence fails to express a proposition and this relieves some of the philosophical disease associated with the liar paradox. Sentence (1′) does not say that sentence (1′) is not true: it doesn't *say* anything. It tries, and fails, to express a proposition.

It is impossible here to expound all of the technical details of Kripke's theory. A more thorough presentation would emphasize relations between Kripke's view and Kleene's three-valued logic, would discuss Kripke's "fixed point theorem" according to which, given certain constraints, there will be a "minimal fixed point," and would detail the way in which Kripke's theory *explains* (grounds) the truth-values of those sentences that have them. Much of the significance of Kripke's work lies precisely in those details. But a couple of points should be noted: to every sentence to which it assigns a value, Kripke's construction assigns the intuitively attractive value. And its failure to assign any value to certain problematic sentences has an important philosophical payoff. Still, there are problems. For example, the sort of construction Kripke proposes fails to assign truth-values to sentences we intuitively expect to have one. Generalizations such as "every true sentence is a true sentence" are ungrounded and left without a truth-value. Attempts to extend Kripke's theory to provide intuitively attractive truth-values for such sentences threaten to undermine the basic intuition of groundedness that gives Kripke's theory much of its force. And though the idea that ungrounded sentences do not express propositions could in principle be eliminated, without that claim the theory's response to the paradox loses much of its philosophical attraction.

Substitutional quantification

In his classic, "Is There a Problem about Substitutional Quantification?" (1976), Kripke establishes a number of important results about substitutional quantification. Quantification (or "generalization"), which can be *existential* or *universal*, involves some schema's being true in at least one case (existential) or in every case (universal). But what is a "case"? This question can introduce the difference between *substitutional* and the perhaps more familiar *objectual* (or sometimes "standard" or "referential") quantification. A true objectual existential quantification requires that there be some *entity* of which the schema is true. A true substitutional existential quantification, by contrast, requires that some *expression* can be substituted for the variable in the schema to produce a true sentence. Important differences between objectual quantification and substitutional quantification arise most clearly when either (1) some names are "empty" (there is nothing in the domain of discourse of which they are the names), or (2) not every entity in the domain of discourse has a name.

Because the truth of generalizations, when they are read substitutionally, can seem not to require the existence of entities in the relevant domain of discourse, substitutional quantification promised to some philosophers an attractive "ontological neutrality." We could say for example that "every even number is divisible by two" without explicitly committing ourselves to the existence of even numbers, if we were so disinclined. But in the early 1970s, papers by J. Wallace and L. Tharp challenged some of

the alleged distinctive value of substitutional quantification. Kripke refuted any suspicion, which some drew from the arguments of Wallace and Tharp, that substitutional quantification is unintelligible or that intelligibly interpreted it reduces to objectual quantification.

Reminding us that substitutional quantification presupposes the notion of a substitution class (the class of items that can be substituted for the variable bound by the substitutional quantifier), Kripke emphasizes that the items in the class must not include the very substitutional quantifier itself. Many of the alleged "paradoxes" surrounding substitutional quantification result from ignoring this requirement. Kripke then shows that it is possible, and in some cases trivial, to give (finitely axiomatized) theories of truth (in Davidson's sense) for languages containing substitutional quantifiers. Theories of truth based on substitutional quantification can, Kripke shows, satisfy Tarski's Convention T. Moreover, Kripke shows that in some cases, a substitutional interpretation of the quantifiers will be equivalent to a referential interpretation. In these cases (which will include all first-order languages without identity), whether the quantifiers are interpreted substitutionally or objectually will make no difference to which formulae are satisfied. But this no more eliminates the difference between substitutional and referential quantification than does the logical equivalence of "P and P" and "P or P" eliminate the difference between disjunction and conjunction.

Although Kripke shows that there is no problem about substitutional quantification, he is skeptical about its role for interpreting natural language. For example, Kripke does not think the viability of substitutional quantification has any bearing on whether the ordinary expressions "there is" or "there exists" typically carry ontological commitment (indeed, he is concerned about the very intelligibility of the "issue" of ontological commitment). Moreover, ordinary existential assertions appear to make no commitment to nameability, as would be required for such quantification to be interpreted substitutionally. At the end of his paper, Kripke draws a series of valuable metaphilosophical morals.

Wittgenstein on following a rule

In his influential *Wittgenstein on Rules and Private Language* (1982), Kripke attempts an exposition of Wittgenstein's so-called private-language argument. Kripke locates that argument earlier in Wittgenstein's *Philosophical Investigations* than was common at the time, earlier, that is, than in the sections that begin with and follow §243. In §201, Wittgenstein says, "this was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule." By starting with this passage, Kripke will emphasize the centrality for the private-language argument of Wittgenstein's considerations on rule following (see WITTGENSTEIN).

Consider the word "plus" or the symbol "+." We use these to express the mathematical function of *addition*. Of course there are infinitely many possible sums: no finite being could ever perform them all. Consider now some sum that we have never performed: Kripke considers 68 + 57. Of course, that sum is 125. But we can imagine a skeptic challenging us. How do we know we're following the same rule for adding as we've used in the past? Why are we so confident that we have always used "+" with the implicit intention that 68 + 57 should turn out to stand for 125? According to Kripke,

the skeptic introduces the possibility that with all of those (finitely many) past uses we really expressed a different function, the "quus" function (or "quaddition"), which is defined to equal x + y so long as x, y < 57, and to equal 5 otherwise. So the skeptic challenges us for some reason to believe that, in order to accord with our past uses of "plus," we should now say "68 + 57 equals 125" rather than "68 + 57 equals 5." If we really did always use "plus" for *quaddition* rather than for *addition*, then in order to do to 68 and 57 what we have in the past done to, say, 3 and 5 to get 8, we should now get 5 as our result.

Kripke admits that the skeptic's hypothesis (that we have always meant *quus* by "plus") is "ridiculous," "fantastic," "bizarre," and "wild." If he proposes it sincerely, the skeptic is surely crazy. But the hypothesis is not logically impossible. If it is false, we should be able to cite some fact about our past usage which establishes that by "plus" I meant *plus* rather than *quus*. The problem is that all candidate facts can seem to fail. Our problem is philosophical: the question is not "do we mean plus by 'plus'?" but "in virtue of what do we mean plus by 'plus'?" If we have no answer to that question then we must take seriously the possibility that meaning is a myth. Of course, in *posing* the paradox we assume that language is meaningful. But we must eventually kick the ladder away: if no fact about us could suffice for our having meant *plus* rather than *quus* in the past (and the paradox is as general as it appears to be), then there can be no fact as to what we mean by anything at any time. Meaning is an illusion.

Much of Kripke's purpose in the book is to develop and sharpen the problem (though he finds material in Wittgenstein to sketch a "skeptical" solution). He deftly deflects several immediate responses. And he devotes a substantial section to discussing a "dispositional" response according to which we mean *plus* rather than *quus* in virtue of having a disposition to perform various calculations in specific ways: we are disposed to give 125, not 5, as the sum of 68 and 57. There are immediate problems such as (1) we might be disposed to perform various calculations erroneously without therefore not meaning *plus* by "plus" and (2) we might have simply *no* disposition with respect to certain additions (if the numbers are too big, for example). But the basic threat to any such response, as Kripke makes clear, is that just because I am in fact disposed to perform various calculations in specific ways does not make it the case that I *should* perform them in that way. If I am performing addition, I *should* derive 125 from 68 and 57, whatever my dispositions might actually be. In Kripke's terminology, the dispositional account of meaning *plus* by "plus" leaves out the *normativity* of meaning. Kripke's discussion has helped make Wittgenstein's rule-following considerations a central issue not only in the philosophy of mind and language, but also in the philosophy of law, where the idea of a rule's having content and normative force, with respect to previously uncontemplated circumstances, is predictably important.

Bibliography of works by Kripke

1959: "A Completeness Theorem in Modal Logic," *Journal of Symbolic Logic* 24, pp. 1–14.

1963a: "Semantical Analysis of Modal Logic I," *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 9, pp. 67–96.

1963b: "Semantical Considerations on Modal Logic," *Acta Philosophica Fennica* 16, pp. 83–94.

1965: "Semantical Analysis of Modal Logic II," in *The New Theory of Models*, ed. J. Addison, L. Henkin, and A. Tarski, Amsterdam: North Holland, pp. 206–20.
1971: "Identity and Necessity," in *Identity and Individuation*, ed. M. Munitz, New York: New York University Press, pp. 135–64. (Also in *Naming, Necessity, and Natural Kinds*, ed. S. Schwartz, Ithaca, NY: Cornell University Press, pp. 66–101.)
1972: "Naming and Necessity," in *Semantics of Natural Language*, ed. D. Davidson and G. Harman, Dordrecht: Reidel, pp. 253–355 and 763–69.
1975: "Outline of a Theory of Truth," *Journal of Philosophy* 72, pp. 690–716. (Also in *Recent Essays on Truth and the Liar Paradox*, ed. R. L. Marin, Oxford: Oxford University Press, 1975, pp. 53–81.)
1976: "Is There a Problem About Substitutional Quantification?," in *Truth and Meaning*, ed. G. Evans and J. McDowell, Oxford: Clarendon Press, pp. 325–419.
1977: "Speaker Reference and Semantic Reference," in *Contemporary Perspectives in the Philosophy of Language*, ed. P. French, T. Uehling, and H. Wettstein, Minneapolis: University of Minnesota Press, pp. 6–27.
1979: "A Puzzle About Belief," in *Meaning and Use*, ed. A. Margalit, Dordrecht: Reidel, pp. 239–83.
1980: *Naming and Necessity*, Cambridge, MA: Harvard University Press, and Oxford: Blackwell Publishers.
1982: *Wittgenstein on Rules and Private Language*, Cambridge, MA: Harvard University Press.
1986: "A Problem in the Theory of Reference: The Linguistic Division of Labor and the Social Character of Naming," in *Philosophy and Culture: Proceedings of the XVIIth World Congress of Philosophy*, Montréal: Éditions du Beffroi, pp. 241–7.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

39

David Lewis (1941–)

ROBERT STALNAKER

Introduction

David Lewis is a philosopher who has written about a wide range of problems in metaphysics and the philosophy of mind and language, including the metaphysics of possible worlds, the analysis of counterfactual conditionals, causation and probability, the problems of universals, of intentionality, of personal identity, the foundations of decision theory, of set theory, of semantics. A distinctive and comprehensive metaphysical theory has emerged from his discussions of philosophical problems: a theory that combines realism about possible worlds with a kind of nominalism, a materialist account of mind, and Humean skepticism about unanalyzed natural necessity. But Lewis's discussions have also yielded conceptual tools that have applications both within and outside of philosophy that are independent of the grand metaphysical scheme, for example, an analysis of common knowledge that has been influential in game theory and theoretical computer science, and work on generalized quantifiers in natural language and on the role of extra-linguistic context in the interpretation of speech that has influenced the development of linguistic semantics.

Lewis studied at Harvard with W. V. Quine and Nelson Goodman, and the influence of those two philosophers is evident in his own philosophical method, in the problems he has focused on, and in the substance of the views he defends. But Lewis developed Quinean and Goodmanian themes with a distinctive twist that takes them in unanticipated directions and that has resulted in a theory that combines features his teachers would applaud with features they would abhor. For Quine and Goodman, the rejection of the analytic/synthetic distinction motivated a holistic philosophical methodology, a method of "reflective equilibrium" that helped to make metaphysics respectable for the heirs of the positivist tradition (see GOODMAN and QUINE). (One consequence of abandoning the two dogmas of empiricism, Quine wrote, was "a blurring of the line between speculative metaphysics and natural science," Quine 1953: 20.) Lewis adopted the holistic method, and accepted the invitation to do metaphysics with a clear conscience, but he defended the analytic/synthetic distinction, and the intelligibility of truth by convention. He followed Goodman in seeking a reductive analysis of counterfactual conditionals, but rejected Goodman's demand for a reduction of the possible to the actual. He adopted Quine's standards for ontological commitment, and for philo-

sophical clarification, but used them to reach very different conclusions about what there is, arguing that Quine's "creatures of darkness" – intensions, propositions, possible worlds – *can* find a place in a world-view that meets the rigorous standards of adequacy that Quine set down. The actual world, according to the metaphysical theory Lewis defends, is much as Quine and Goodman thought. Their only mistake was to think that the actual world is the only world there is.

The emphasis in this exposition will be on the general metaphysical framework that provides the context for Lewis's many constructive philosophical analyses. I will begin with some general remarks about philosophical method and metaphysics in the next section, and after that discuss Lewis's modal realism and finally his Humean account of counterfactuals, laws, and causation.

Method and metaphysics

During the first half of the twentieth century, the word "metaphysics" had mostly a pejorative use within the analytic philosophical tradition. The logical empiricists taught that metaphysics was the result of equivocation between questions about meaning, which called for a decision about what linguistic framework to use and questions that arise within the context of an accepted framework. But Quine noted that the methods used within the scientific framework for deciding which theoretical claims were true were not very different from the methods used to make the practical decisions about what language forms to adopt. In both cases, one chose the theory or framework that did the best job of making sense of one's experience. He argued that the line between internal and external questions, and between decisions that constituted linguistic stipulation and decisions that constituted empirical judgments was arbitrary. If decisions about what general framework to theorize in are not separable from judgments about what is true, then there is room for metaphysics after all. "The quest of a simplest clearest overall pattern of canonical notation," Quine wrote, "is not to be distinguished from a quest of ultimate categories, a limning of the most general traits of reality" (Quine 1960: 161).

Lewis's account of his philosophical method follows that of Quine and Goodman closely. We begin with a collection of opinions. "Some are commonsensical, some are sophisticated; some are particular, some general. . . . A reasonable goal for a philosopher is to bring them into equilibrium," Lewis 1983b: x) And like Quine, Lewis emphasizes that the method of reflective equilibrium should not be taken to have relativist or anti-realist consequences. Philosophy may be a matter of opinion, but some opinions, even some that are in some philosopher's reflective equilibrium, may nevertheless be false. But unlike Quine and Goodman, Lewis did not tie his epistemological holism to the rejection of the analytic/synthetic distinction. His first major philosophical project responded to Quine's critique of this distinction, and of truth by convention. Lewis accepted the terms of Quine's demand for an analysis: one must break out of the tight circle of concepts (synonymy, semantic rule, meaning, etc.), and explain what it is to be an analytic truth in terms of the dispositions and behavior of language users. But he argued that this could be done with the help of a general analysis of the notion of a *convention*, and a distinction between two different notions of language: language as defined by a set of syntactic and semantic rules and language as defined by a popula-

tion of speakers. The definition of an abstract language simply stipulates that the language is constituted by certain semantic rules that determine a class of analytic truths. The work is done in explaining, in terms of an analysis of convention, what it is about the behavior, expectations, dispositions of a given population of speakers for a language defined in this way to be the language spoken by that population.

There are two ways in which Lewis's account of analyticity, even if fully adequate on its own terms, will fail to satisfy an unreconstructed Quinean. First, conventions are explained in terms of intentions, beliefs, and knowledge, and so an explanation of semantic notions such as meaning and analyticity in terms of convention would not be an explanation that solved the problem of intentionality. Quine thought that to the extent that mentalistic intentional notions such as belief and intention could be explained at all, they would be explained in terms of the intentionality of language – believing, for example, in terms of holding true – and so he would not be satisfied with an explanation of semantic notions that took belief and intention for granted. In this regard, Lewis is like H. P. Grice, separating problems about linguistic meaning from the more general problem of intentionality, and taking the intentionality of thought as more basic. But Lewis, Quine, and Grice would all agree that whichever comes first, an adequate account of linguistic and mental intentionality must ultimately explain them in materialistically acceptable terms.

A second way in which the account will disappoint a Quinean is perhaps the more significant one. As Lewis emphasized, his account of analyticity makes reference to possible worlds, and so does not provide an informative analysis of one of the notions – necessity – that Quine would have put in his tight circle of problematic concepts. But Lewis argued that the metaphysical notions of necessity and possibility do not belong in this circle, since they are not semantic notions. Analytic truths are necessary because they express propositions that are necessary. The account of the conventions of language explain why the sentences used by the members of some population express the propositions they express, but the necessity or contingency of the propositions themselves has nothing to do with convention, or with language.

Though he defends analyticity, Lewis does not assume that speakers are authoritative about the conventions of their own language, so about the analytic truths. Even if there is a sharp line between truths of meaning and truths of fact, there is no sharp line between linguistic intuition and beliefs about substantive theory. "Our 'intuitions' are simply opinions, and our philosophical theories are the same" (Lewis 1983b: x). We can draw the line between analytic and synthetic, but the decision about where we draw it, like our other decisions about what to believe, is a part of a judgment about the global theory that, all things considered, best makes sense of our experience.

Modal realism

Possible worlds have played a prominent role in Lewis's philosophical analyses from the beginning. Being a good Quinean, Lewis recognized an obligation either to admit them into his ontology, or to reduce them to something else. And if they are to be accepted, it should be clear what kind of thing they are. "We ought to believe in other possible worlds and individuals," he argues, "because systematic philosophy goes more smoothly in many ways if we do" (Lewis 1986b: 354). Lewis makes no attempt to mini-

mize the counterintuitive character of the ontological commitment he is prepared to make; possible worlds, as he uses the term, are concrete particulars: other things of the same kind as the universe of which we are a part. Merely possible highways in lands that will never be actual are made of concrete that is just as real as that used to make the actual highways on which we drive. Just as people who live at other times and places in the actual world are as real as we are, so, according to Lewis's modal realism, are the non-actual people who inhabit other possible worlds. Their nonactuality consists in the fact that they are spatiotemporally disconnected from us. Lewis grants – in fact emphasizes – that the belief in a plurality of parallel universes conflicts sharply with common opinion, and since he takes common opinion seriously, he acknowledges that this is a serious cost to be balanced against the benefits of this metaphysical theory. But while he takes the "incredulous stare" that is a common response to this theory to reflect a formidable objection, he argues that more theoretical arguments against modal realism fail, as do attempts to analyze possible worlds away, or to give a more innocent explanation of what they are. In the end, he judges that the cost of offending common opinion is outweighed by the many benefits that modal realism brings. So the strategy for defending modal realism combines an exposition of the many benefits of the possible worlds framework, responses to theoretical arguments against modal realism, and arguments against attempts to get those benefits without the counterintuitive commitment.

It is useful to divide the doctrine of modal realism into a semantic and a metaphysical component. First, there is the metaphysical thesis that there is a large plurality of parallel universes, where a single universe consists of everything that is spatiotemporal related to anything in it. Second, there are the semantic analyses that relate this plurality of worlds to the many modal, epistemic, and intentional concepts whose clarification provide the benefits of modal realism. As an example of a thesis that belongs to the second component, consider the analysis of possibility as truth in some possible world. Lewis emphasizes that the theory must be evaluated as a package, and he would agree that each component would loose all plausibility without the other. On the assumption that the metaphysical thesis is false – that common sense is right that our universe is the only one – the semantic analysis of possibility has no plausibility, since on that assumption the possible collapses into the actual. On the other hand, if we look at the metaphysical claim in isolation from the semantic analyses, it looks like an extravagant and gratuitous empirical hypothesis. Why should one believe in all these other universes? Lewis's answer – that systematic philosophy goes more smoothly if we do – has force only when the metaphysical hypothesis is combined with the semantic analyses that connect the hypothesis with the phenomena that systematic philosophy seeks to explain.

The possible worlds framework promises to clarify not only *de dicto* modal claims, such as that it is necessary that all bachelors are unmarried, but also *de re* modal claims such as that no bachelor is *essentially* unmarried. Modal realism uses *counterpart theory* to analyze claims about the modal properties of things. As with the general modal realist thesis, we can distinguish a metaphysical and a semantic component of Lewis's counterpart theory. There is the metaphysical claim that individuals exist in only one possible world, and the semantic claim that *de re* modal properties should be analyzed in something like the following way: an individual has the property of being possibly

F if and only if it has a counterpart that has the property of being *F*, where the counterpart relation is a contextually determined relation of similarity in relevant respects. As with the general thesis, Lewis would emphasize that the semantic and metaphysical parts of the package must be evaluated together.

The metaphysical doctrine has been criticized (for example, by Alvin Plantinga and Nathan Salmon) on the ground that it has the implausible consequence that all properties are essential properties; but this criticism simply assumes that Lewis's semantic analysis of what it is to have a property essentially is mistaken. The semantic thesis has been criticized (for example by Saul Kripke) on the ground that it has the consequence that when we say that Humphrey might have won the election, we are not really talking about *Humphrey*. But Lewis rightly insists that, on the counterpart analysis, it is Humphrey himself who has, in the actual world, the property of being a possible winner. It is just that he has this property in virtue of his resemblance to someone else who (in another possible world) has the property of being a winner. The counterpart semantics may be more complex and less straightforward than the standard analysis, but given the metaphysical thesis, it gives a better account of our modal beliefs. And if one accepts the general doctrine that other possible worlds are parallel universes, it seems most reasonable to think that no one can inhabit more than one of them. Whatever one's verdict about the plausibility of modal realism as a whole, it seems clear that counterpart theory belongs in the package.

Both modal realism's central metaphysical thesis and its semantic analyses of necessity, possibility, and other modal notions conflict with unreflective common opinion. It is not only that it strains credibility to hypothesize that there is a vast plurality of parallel universes, it also seems counterintuitive to many people to claim that our opinions about what might or would have happened are opinions about the existence of such parallel universes. As noted above, Lewis grants that modal realism conflicts with unreflective common opinion, and that this conflict is a strike against the theory, but he argues that the cost is outweighed by the benefits. Since he agrees that if some alternative account could provide the benefits without the cost, modal realism would not be defensible, it is an important part of its defense to criticize attempts to reconcile the explanations that the possible worlds framework provides with a more modest account of what possible worlds are.

"Ersatz modal realism" is Lewis's label for the attempt to get the benefits of modal realism without the costs by explaining possible worlds as something other than parallel universes. Most of his critical discussion of this project is devoted to attempts to reduce possible worlds to some kind of linguistic object: state descriptions, maximal consistent sets of sentences, complete novels. This is a common and seductive strategy for explaining what possible worlds are, but there is a lot wrong with it, as Lewis's criticisms bring out. The most serious problem is that this kind of explanation seems to foreclose one of the most important uses of possible worlds: to represent the contents of speech acts and propositional attitudes. If sentences, or sets of them, are to represent possible worlds adequately, they must be interpreted sentences – sentences with their truth conditions. We will have a serious circularity if we try to combine this kind of explanation of possible worlds with an explanation of the truth conditions of a sentence in terms of the possible circumstances, or possible worlds, in which the sentence would be true. There are, however, philosophical accounts of possible worlds that agree

with Lewis that possible worlds are non-linguistic things, suitable for representing truth-conditional content, while disagreeing with the thesis that possible worlds are something like other universes parallel to our own.

According to the simplest and most straightforward attempt to explain what possible worlds are in a way that is compatible with actualism (a thesis that common opinion might regard as trivially true: that what actually exists is all there is), possible worlds (or less misleadingly, possible states of the world) are a kind of property: ways the world might be, or might have been. This is obviously not a reduction of possible worlds to something else: it is intended simply as a characterization of the kind of thing that a possible world is. To call possible worlds properties is to say two things about them: first, they are things that are, or may be, instantiated. Second, they are the kind of thing that is (at least prima facie) independent of language and thought. A property of something (such as the property of being the first child born in the twenty-first century) is different both from the thing (if any) that has the property and from a thought or a predicate that expresses the property. The significance of this characterization is that it provides us with a way to reconcile a commitment to the existence of possible worlds, construed as non-mental entities, with the apparently contradictory thesis that the actual world is the only world there is.

Anyone who takes literally the claim that there are possible worlds has to respond to this prima-facie paradox: unrealized possibilities – counterfactual situations – are situations that turned out not to exist. How can there *be* situations, or worlds, that don't exist? Any response to this problem will make a distinction between a sense in which non-actual possible worlds exist, and a sense in which they do not. Lewis's strategy is to distinguish two different scopes for the quantifier. Quantifiers are often restricted to some contextually determined subdomain of all there is, and one very general restriction, according to Lewis, is to the domain of things that inhabit the actual world. When we say that there are no talking donkeys, we normally mean that there are no *actual* talking donkeys. But there is also an unrestricted quantifier, which ranges over absolutely everything there is. Common opinion may not distinguish what exists from what actually exists, but Lewis would say that the distinction is implicit in their modal discourse. The actualist response to this puzzle makes the distinction, not in terms of a difference of domain, but in terms of an ambiguity in the terms "possible world"and "actual world." Just as we can distinguish the property of being the first child born in the twenty-first century from that child, so we can distinguish the property of being a universe of a certain kind from a universe that is of that kind. According to the actualist, there are (and *actually* are) many ways the world might have been, but there is only one world that is one of those ways.

This construal of possible worlds as properties allows us many of the benefits of the possible worlds framework (for example, the formal semantic analysis of modal and epistemic notions, the clarification of counterfactuals and causal and temporal structures, the representation of probability as a measure on state spaces, the representation of mental and linguistic content, and of speech contexts) without either denying the ontological commitment to possible states of the world, or challenging pretheoretical common opinion. It does not produce incredulous stares to say that there are many ways the world might have been. To see why Lewis resists this actualist interpretation of possible worlds we need to consider another one of his metaphysical priorities that

has its root in the Quine–Goodman legacy: a penchant for nominalism (see GOODMAN and QUINE).

One of the benefits of modal realism, according to Lewis, is that it provides us with an analysis of properties: of what properties are, and what it is to have a property. Properties, according to Lewis's modal realism, are just sets, and to have a property is just to be one of its members. The domain of all possibilia provides an answer to the standard objection to the identification of properties with their extensions, an answer that those with only the impoverished domain of the actual things cannot avail themselves of. Distinct properties can have the same extension in the actual world, but they are distinguished by the difference in their extensions in other possible worlds. Even if in the actual world, all and only creatures with a kidney are creatures with a heart, the two properties are distinguished by the fact that there are possible creatures with one property, but not the other. So modal realism offers the virtues of a simple extensionalist account of properties without the defects of actualist versions of that account.

Lewis recognizes that an adequate theory of properties needs to distinguish between different kinds of properties. Some properties (sets) and relations (sets of n-tuples) are *natural* or *fundamental*. Among the fundamental relations, some are *spatiotemporal* relations. These are primitive distinctions of Lewis's theory, but he argues that they are distinctions that any plausible metaphysical theory must make. With just these primitive concepts for classifying properties and relations, Lewis suggests, we can give a full characterization of the logical space of possible worlds while continuing to maintain that properties are nothing but sets.

World-mates (inhabitants of the same possible world) are individuals that stand in spatiotemporal relations with each other. A possible world is fully characterized by specifying a set of world-mates, and by saying which fundamental properties they have, and how they are related by the fundamental relations. All the properties and relations of the things in any world *supervene* on the fundamental properties and relations of those things: possible worlds that are indiscernible from each other with respect to fundamental properties and relations are identical. This a priori supervenience claim is not substantive, since the fundamental properties are just those that are necessary to give a complete characterization of a possible world. Substantive (and contingent) metaphysical hypotheses can be stated as theses about what the fundamental properties of the actual world are. So, for example, *materialism* is explained as the thesis that only physical properties and relations are fundamental. (That is, materialism is true of possible world *w* if the fundamental properties and relations of things in *w* are all physical.) The thesis of *Humean supervenience,* which we will discuss in the next section, is the thesis that only intrinsic properties and spatiotemporal relations are fundamental.

The theory of properties as sets is a crucial part of Lewis's modal realism, and unlike many of the fruits of the possible worlds framework, this analysis cannot be reconciled with the actualist interpretation of possible worlds. Lewis's theory can, of course, make the distinction to which the actualist appeals between properties and their instances – it is just the distinction between sets and their members – but it will be no help in avoiding a commitment, not just to ways the world might be, but to worlds that are those ways. For if properties are sets, and if possible states of the world are identified with maximal properties that the world might have, then a possible state of the world is a

unit set with a world that is in that state as its member. According to the actualist metaphysics, there is only one thing to be the member of such a set, and so if possible states of the world are properties, and Lewis is right about what properties are, there is only one possible state of the world. So while actualists can avail themselves of many of the benefits of the framework of possible worlds, they will have to forego Lewis's elegant reductive account of properties.

Counterfactuals and causation

Lewis's second book undertook to give a reductive analysis of counterfactual conditionals, a project motivated by the same Humean skepticism about natural necessity that motivated Nelson Goodman to try to give such an analysis more than twenty years earlier. For Goodman, the crux of the problem was the modal character of counterfactuals: they seemed to be about unrealized possibilities. The task was to explain the possible in terms of the actual. Lewis, of course, had no problem with non-actual possibilities, and took counterfactuals at face value as statements about counterfactual possible worlds. But counterfactuals (and statements about cause and effect, dispositions and propensities, dependency and chance) are for the most part contingent statements. One has to explain how a statement about counterfactual possibilities can be contingently true or false in the actual world. For such statements to be contingent, the counterfactual worlds that are relevant to the evaluation of a conditional must be determined by their relation to the actual world.

Lewis's formal semantics gives truth conditions for conditionals in terms of a three-place *comparative similarity relation* on possible worlds (world x is more similar to world w than y is to w). The rough idea of the analysis is that a conditional, "if A, then C" is true (in a possible world w) if and only if C is true in those possible worlds in which A is true that are most similar to w. This first approximation is not quite right, since if there is an infinite sequence of ever more similar worlds in which A is true, there will be no closest such possible worlds. To allow for this case, Lewis's favored analysis is as follows: "If A, then C" is true in w if and only if some world in which A&C is true is closer to w than any world in which A&~C is true. This analysis provides an abstract formal semantics for counterfactual conditionals, but we don't have a reductive analysis until we have explained the relevant respects of similarity. The semantic analysis is just the first step of a larger project, a defense of the doctrine that Lewis labeled "Humean supervenience." The project is motivated by a Humean skepticism about real relations between "distinct existences."

For the Humean, spatiotemporal relations (such as contiguity) are acceptable, as are logical relations, or relations of ideas. Relations of resemblance between things are acceptable, so long as the respects of resemblance are spelled out, since they are explicable in term of the sharing of specified properties. But causal relations, and others in the same family, must be analyzed in terms of global regularities, with the help of relations of the unproblematic kind. If the respects of similarity between possible worlds that are relevant to the interpretation of counterfactuals can be specified, Lewis's analysis will yield an account of counterfactuals that should satisfy a Humean, and so an account that permits the Humean to use counterfactuals to analyze relations of causation and causal dependence and independence.

Lewis's Humean project has the following separable components: (1) an abstract semantic analysis of conditionals in terms of comparative similarity; (2) an explanation of the respects of similarity that are appropriate for interpreting of the kind of conditionals that are relevant to the analysis of causal relations. Since Lewis's response to this problem appeals to laws of nature, he needs (3) an account of laws of nature in terms of global patterns of particular fact, and finally, (4) an analysis of causation in terms of counterfactuals. This is an ambitious agenda. Some parts have been carried out in detail and with precision; in other cases, there are only sketchy suggestions about the kind of account that should be given. And some parts of the project are ongoing.

A conditional is true if the consequent is true in the possible world in which the antecedent is true that is most similar, in relevant respects, to the actual world. But what are the relevant respects? One might be tempted to appeal to an intuitive notion of overall similarity. Lewis notes that we do make and understand judgments of overall similarity between complex object such as cities, and we do have intuitions about which possible worlds are more and less alike. A general impressionistic notion of similarity would be both vague and context-dependent, but as Lewis notes, counterfactuals are both vague and context-dependent. There would, however, be at least two problems with relying on such a notion of similarity. First, an impressionistic notion of similarity would be suitable for the project of Humean reduction, since judgments of similarity between worlds might be based in part on comparison of unanalyzed facts about causal relations. But second, in any case there are counterexamples that show that overall similarity is not the right relation. It seems intuitively clear that small events can have large consequences. If Oswald had missed Kennedy in 1963, the course of American politics between then and now probably would have been quite different. But isn't a possible world in which Oswald misses, but someone else succeeds, and the course of American politics proceeds much as it actually did much more similar, overall, to the actual world? If certain conspiracy theorists are right, and there were backup assassins ready to act if Oswald failed, then it might be true that if Oswald hadn't killed Kennedy, someone else would have, but we don't want an analysis of counterfactuals to ensure that such conspiracy theories are true.

One might be tempted to build a temporal asymmetry into the account of comparative similarity that is relevant to the interpretation of counterfactuals: perhaps similarity of earlier parts of history should have much greater weight than similarity of later times. But to do this would be to explain the temporal asymmetry of causal and counterfactual dependence as a consequence of convention and not as a fact about the world. Lewis's aim was to define a temporally neutral notion of comparative similarity between possible worlds, and use it to explain how temporal asymmetries in the pattern of facts in the actual world results in a *de facto* asymmetry of counterfactual dependence.

Lewis's account of the relevant respects of comparative similarity between worlds gives highest priority to avoiding large and widespread violations of laws of nature. The second priority is to maximize *exact* agreement of particular fact. Small and local violations of laws of nature are permissible to achieve the second priority, and in a deterministic world, such "small miracles" will always be required. *Approximate* agreement of fact counts for very little: deviations from the laws, even small ones, in order to increase approximate similarity of fact are not permitted, and possible worlds that agree

exactly for a period of time are more similar than worlds that agree only approximately, but over a much longer period of time. There is no attempt to make this account of comparative similarity precise, but Lewis argues that it helps to explain some temporal asymmetries, and that it points to the kind of explanation that can vindicate a reductive account of counterfactuals.

A Humean cannot, of course, rest with an appeal to an unanalyzed notion of law of nature. Here Lewis endorses an idea of Frank Ramsey's: that the laws of nature (in a given possible world) are the factual regularities that are consequences of the simplest and strongest systematization of the truths of that world. The criteria for evaluating systems of truths remain to be explained, but Lewis argues that so long as the relevant criteria of strength and simplicity are non-contingent, this will be an account of law of nature that meets the standards of Humean supervenience.

Counterfactuals, once explained in terms of resemblance of the facts and regularities of the possible worlds, are then available to the Humean for an analysis of causation. The first step is to define counterfactual dependence: one truth B counterfactually depends on another A if B would not have been true if A had not been true. If *c* and *e* are distinct events that occur, it seems a good first approximation to say that *c* is a cause of *e* if and only if both occur, and *e* would not have occurred if *c* had not. This proposal would account for many of the features of causation that create problems for a simple regularity analysis. Cause can be distinguished from effect without making explicit appeal to temporal order, and events that are regularly connected because one causes the other can be distinguished from events that are connected because they are each effects of a common cause.

But cases of preemptive causes show that one cannot, in general, identify causation with counterfactual dependence. Suppose the hit man was successful, but if he had missed, another was waiting in the wings to do the job. The victim's death was caused by the hit man's action, but, because of the backup potential cause, was not counterfactually dependent on it. Lewis's first strategy for accommodating preemptive causation was to define the causal relation as the transitive closure of the relation of counterfactual dependence (between distinct events). The death does not depend counterfactually on the shooting, but there will be intermediate events which are dependent on the shooting, and on which the effect is dependent. This move accounts for some cases of preemption, but not for all. A second strategy for dealing with preemption cases argues that while the man would still have died if the backup assassin had done the job, he would have died a different death, and so despite the preemption, the event that was the effect was still counterfactually dependent on the actual assassin's act. But it is difficult to find and motivate an account of the modal properties of events that will explain all cases of preemption in this way without intuitively implausible consequences. Even taking the resources of counterfactual conditionals for granted, the analysis of causation has proved to be a surprisingly recalcitrant problem. This is now a lively area of ongoing research.

Conclusion

Lewis's metaphysical framework, and his philosophical method, provide a rich context for the clarification of philosophical problems, the articulation and defense of philo-

sophical theses, and the formulation of constructive conceptual analyses. He has formulated and defended a materialist theory of mind, with accounts both of intentional states such as belief, and sensory states such as pain. In the context of the defense of Humean supervenience, he explored the relation between objective and subjective probability – chance and degree of belief – and between subjective probability and conditional propositions. And he provides a foundation for causal decision theory. Building on his early work on convention, he developed a foundation for semantics and pragmatics that clarifies the relations between speech and thought, and that also makes substantive contributions to compositional semantic theories for natural languages. Modal realism's use of set theory motivated an exploration of the foundations of set theory itself that clarifies the relation between mereology (the theory of parts and wholes) and set theory.

Even though Lewis's general metaphysical theory has a coherence and unity that tie the different parts together, many of his constructive analyses are separable from the system that provides the context for their development. This is appropriate, given Lewis's pragmatic cost–benefit methodology: he recognizes that others with different priorities may not be prepared to swallow his system whole. Those who reject modal realism or Humean supervenience will still find much to learn and to adopt from his philosophical work. And even those who are skeptical about this metaphysical theory can appreciate the power of a system that has generated so many clarifying philosophical analyses.

Bibliography

Works by Lewis

1969: *Convention*, Cambridge, MA: Harvard University Press.
1976: *Counterfactuals*, Oxford: Blackwell Publishers.
1983a: "New Work for a Theory of Universals," *Australasian Journal of Philosophy* 61, pp. 343–77.
1983b: *Philosophical Papers*, vol. I, Oxford: Oxford University Press.
1986a: *On the Plurality of Worlds*, Oxford: Blackwell Publishers.
1986b: *Philosophical Papers*, vol. II, Oxford, Oxford University Press.

Works by other authors

Quine, W. V. (1953) *From a Logical Point of View*, Cambridge, MA: Harvard University Press.
——(1960) *Word and Object*, Cambridge, MA: MIT Press.

Blackwell Companions to Philosophy: A Companion to Analytic Philosophy
Edited by A. P. Martinich, David Sosa

Index

Note: Page references in bold type indicate the main treatment of the philosopher.