Reflexivity: A Source Book in Self-reference

edited by Steven James Bartlett

About the 2015 online edition of this work

Reflexivity: A Source Book in Self-reference was originally published in 1992 by North-Holland, an imprint of Elsevier Science Publishers B. V. The book is now out-of-print. Elsevier has consequently granted to the Editor a reversion of rights to the book.

To make this book available at no cost to its readers, the Editor has chosen to re-issue the book as a free open access publication under the terms of the Creative Commons **Attribution-NonCommercial-NoDerivs** license, which allows anyone to distribute this work without changes to its content, provided that both the author and the original URL from which this work was obtained are mentioned, that the contents of this work are not used for commercial purposes or profit, and that this work will not be used without the author's or his executor's permission in derivative works (i.e., you may not alter, transform, or build upon this work without such permission). The full legal statement of this license may be found at

http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode

A note on pagination: Page numbers enclosed in brackets that appear toward the bottom of the page are sequentially numbered and refer to the page numbering of the originally printed book. Page numbers not enclosed in brackets are those of the original published sources of the individual papers.

REFLEXIVITY

A Source-Book in Self-Reference

Edited and with an Introduction by

STEVEN J. BARTLETT

Research Professor of Philosophy
Oregon State University
and
Visiting Scholar in Psychology and Philosophy
Willamette University

1992

NORTH-HOLLAND
AMSTERDAM • LONDON • NEW YORK • TOKYO

ELSEVIER SCIENCE PUBLISHERS B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

Distributors for the United States and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY INC.
655 Avenue of the Americas
New York, N.Y. 10010, U.S.A.

Library of Congress Cataloging-in-Publication Data

Reflexivity : a source-book in self-reference / edited and with an introduction by Steven J. Bartlett.
p. cm.
Includes bibliographical references and index.
ISBN 0-444-89092-0
1. Reference (Philosophy) 2. Self-knowledge, Theory of.
I. Bartlett, Steven J.
B105.R25R44 1992
110--dc20 92-919
CIP

ISBN: 0 444 89092 0

Printed in The Netherlands.

Dedicated to my father

Paul Alexander Bartlett

Whose gentle and caring spirit
and love of beauty touched
those who knew him.

Companion on many hacienda trips,
a very special friend,
and now,
greatly missed.

ACKNOWLEDGEMENTS

I would like to express my special thanks to Henry W. Johnstone, Jr., for his faithful friendship and encouragement in this project.

Thanks are due to the following authors, journals, and conference proceedings for their permission to reprint papers found in this volume: (The order follows that of the table of contents.)

W. V. Quine, "Paradox," reprinted from *Scientific American*, Vol. 20, No. 4, 1962, pp. 84-96.

Paul Weiss, "The Theory of Types," reprinted from the *Review of Metaphysics*, Vol. 4, 1951, pp. 338-348.

John Myhill, "A System Which Can Define its Own Truth," reprinted from *Fundamenta Mathematicae*, Vol. 37, 1950, pp. 190-192.

Gilbert Ryle, "Heterologicality," reprinted from *Analysis*, Vol. 11, 1950-51, pp. 61-69.

Jørgen Jørgensen, "Some Reflections on Reflexivity," reprinted from *Mind*, Vol. 62, 1953, pp. 289-300.

R. M. Martin, "On Non-translational Semantics," reprinted from *Actes du XIème Congrès International de Philosophy*, Vol. 5, 1953, pp. 132-138.

Raymond M. Smullyan, "Languages in Which Self Reference Is Possible," reprinted from the *Journal of Symbolic Logic*, Vol. 22, 1957, pp. 55-67. Copyright © 1957 by the Association for Symbolic Logic. This reproduction by special permission.

A. N. Prior, "On a Family of Paradoxes," reprinted from the *Notre Dame Journal of Formal Logic*, Volume II, Number 1 (January 1961) on pages 16-32, with permission.

Nicholas Rescher, "A Note on Self-Referential Statements," reprinted from the *Notre Dame Journal of Formal Logic*, Volume 5, Number 3 (July 1964) on pages 218-220, with permission.

Robert L. Martin, "Toward a Solution to the Liar Paradox," reprinted from the *Philosophical Review*, Vol. 76, 1967, pp. 279-311.

Bas C. van Fraassen, "Presupposition, Implication, and Self-Reference," reprinted from the *Journal of Philosophy*, Vol. 65, 1968, pp. 136-152.

D. J. O'Connor, "Pragmatic Paradoxes," reprinted from *Mind*, Vol. 57, 1948, pp. 358-359.

L. Jonathan Cohen, "Mr. O'Connor's 'Pragmatic Paradoxes,'" reprinted from *Mind*, Vol. 59, 1950, pp. 85-87.

Peter Alexander, "Pragmatic Paradoxes," reprinted from *Mind*, Vol. 59, 1950, pp. 536-538.

Austin Duncan-Jones, "Fugitive Propositions," reprinted from *Analysis*, Vol. 10, 1949-1950, pp. 21-23.

D. J. O'Connor, "Pragmatic Paradoxes and Fugitive Propositions," reprinted from *Mind*, Vol. 60, 1951, pp. 536-538.

C. K. Grant, "Pragmatic Implication," reprinted from *Philosophy*, Vol. 33, 1958, pp. 303-324.

W. D. Hart, "On Self-Reference," reprinted from the *Philosophical Review*, Vol. 79, 1970, pp. 523-528.

Frederic B. Fitch, "Self-Reference in Philosophy," reprinted from *Mind*, Vol. 55, 1946, pp. 64-73.

Frederic B. Fitch, "Universal Metalanguages for Philosophy," reprinted from the *Review of Metaphysics*, Vol. 17, 1963-64, pp. 396-402.

Steven J. Bartlett, "The Idea of a Metalogic of Reference," reprinted from *Methodology and Science*, Vol. 9, 1976, pp. 85-92.

Steven J. Bartlett, "Referential Consistency as a Criterion of Meaning," reprinted from *Synthese*, Vol. 52, 1982, pp. 267-282.

J. McCarthy, "First Order Theories of Individual Concepts and Propositions," reprinted from *Machine Intelligence*, Vol. 9, 1979, pp. 129-147.

Hector J. Levesque, "Foundations of a Functional Approach to Knowledge Representation," reprinted from *Artificial Intelligence*, Vol. 23, 1984, pp. 155-212.

Kurt Konolige, "A Computational Theory of Belief Introspection," reprinted with permission from International Joint Conferences on Artificial Intelligence, Inc., *Proceedings International Joint Conference on Artificial Intelligence, 1985*, pp. 502-508; copies of this and other IJCAI Proceedings are available from Morgan Kaufmann Publishers, Inc., 2929 Campus Drive, San Mateo, CA 94403 U.S.A.

Donald Perlis, "Languages with Self-Reference, I: Foundations," reprinted from *Artificial Intelligence*, Vol. 25, 1985: pp. 301-322.

Donald Perlis: "Languages with Self-Reference, II: Knowledge, Belief, and Modality," reprinted from *Artificial Intelligence*, Vol. 34, 1988, pp. 179-212.

Paul Weiss, "Cosmic Necessities," reprinted from the *Review of Metaphysics*, Vol. 4, 1951, pp. 359-375.

Robert J. Richman, "On the Self-Reference of a Meaning Theory," reprinted from *Philosophical Studies*, Vol. 4, 1953, pp. 69-72.

Henry W. Johnstone, Jr., "Argumentation and Inconsistency," reprinted from the *Revue Internationale de Philosophie*, Vol. 15, 1961, pp. 353-365.

Alf Ross, "On Self-Reference and a Puzzle in Constitutional Law," reprinted from *Mind*, Vol. 78, 1961, pp. 1-24.

Joseph M. Boyle, Jr., "Self-Referential Inconsistency, Inevitable Falsity and Metaphysical Argumentation," reprinted from *Metaphilosophy*, Vol. 3, 1972, pp. 25-42.

• • •

I would like to express my thanks to Drs. Arjen Sevenster, Elsevier's Senior Editor of Mathematics, Computer Science, and Cognitive Science, for his interest in this project, his helpful suggestions, and faithful collaboration. I wish also to thank Miss Titia Kraaij, Elsevier's Technical Editor, for her competent and conscientious supervision of the production of this volume.

CONTENTS

INTRODUCTION

The Role of Reflexivity in Understanding Human Understanding
Steven J. Bartlett 3

PART I: SEMANTICAL SELF-REFERENCE

Paradox
W. V. Quine 21

The Theory of Types
Paul Weiss 37

A System Which Can Define its Own Truth
John Myhill 49

Heterologicality
Gilbert Ryle 53

Some Reflections on Reflexivity
Jørgen Jørgensen 63

On Non-translational Semantics
R. M. Martin 75

Languages in Which Self Reference Is Possible
Raymond M. Smullyan 83

On a Family of Paradoxes
A. N. Prior 97

A Note on Self-Referential Statements
Nicholas Rescher 115

Toward a Solution to the Liar Paradox
Robert L. Martin 119

Presupposition, Implication, and Self-Reference
Bas C. van Fraassen 153

PART II: PRAGMATICAL SELF-REFERENCE

Pragmatic Paradoxes
D. J. O'Connor 173

Mr. O'Connor's "Pragmatic Paradoxes"
L. Jonathan Cohen 175

Pragmatic Paradoxes
Peter Alexander 179

Fugitive Propositions
Austin Duncan-Jones 183

Pragmatic Paradoxes and Fugitive Propositions
D. J. O'Connor 187

Pragmatic Implication
C. K. Grant 191

On Self-Reference
W. D. Hart 213

PART III: METALOGICAL SELF-REFERENCE

Self-Reference in Philosophy
Frederic B. Fitch 221

Universal Metalanguages for Philosophy
Frederic B. Fitch 231

The Idea of a Metalogic of Reference
Steven J. Bartlett 239

Referential Consistency as a Criterion of Meaning
Steven J. Bartlett 247

PART IV: COMPUTATIONAL SELF-REFERENCE

First Order Theories of Individual Concepts and Propositions
J. McCarthy 265

Foundations of a Functional Approach to Knowledge Representation
Hector J. Levesque 285

A Computational Theory of Belief Introspection
Kurt Konolige 343

Languages with Self-Reference, I: Foundations
Donald Perlis 363

Languages with Self-Reference, II: Knowledge, Belief, and Modality
Donald Perlis 385

PART V: SELF-REFERENTIAL ARGUMENTATION

Cosmic Necessities
Paul Weiss 421

On the Self-Reference of a Meaning Theory
Robert J. Richman 439

Argumentation and Inconsistency
Henry W. Johnstone, Jr. 443

On Self-Reference and a Puzzle in Constitutional Law
Alf Ross 457

Self-Referential Inconsistency, Inevitable Falsity and Metaphysical Argumentation
Joseph M. Boyle, Jr. 481

AUTHOR INDEX 499

SUBJECT INDEX 503

Publisher's note

The articles in this collection were reproduced in facsimile. The page numbers in the Table of Contents refer to the pagination which we added during the preparation of this volume. These numbers are given in square brackets at the bottom of each page.

The original lay-out of two articles (pp. [21-26] and [243-362]) had to be adjusted to conform to the format of the present volume.

INTRODUCTION

THE ROLE OF REFLEXIVITY IN UNDERSTANDING HUMAN UNDERSTANDING

STEVEN J. BARTLETT

Philosophy is reflective. The philosophizing mind never simply thinks about an object, it always, while thinking about any object, thinks also about its own thought about the object. Philosophy may thus be called thought of the second degree.

Philosophy . . . has this peculiarity that reflection upon it is part of itself. . . . [T]he theory of philosophy is itself a problem for philosophy; and not only a possible problem, but an inevitable one.

— R. G. Collingwood[1]

THE INTERNAL LIMITATIONS OF HUMAN UNDERSTANDING

We carry, unavoidably, the limits of our understanding with us. We are perpetually confined within the horizons of our conceptual structure. When this structure grows or expands, the breadth of our comprehensions enlarges, but we are forever barred from the wished-for glimpse beyond its boundaries, no matter how hard we try, no matter how much credence we invest in the substance of our learning and mist of speculation.

The limitations in view here are not due to the mere finitude of our understanding of ourselves and of the world in which we live. They are limitations that come automatically and necessarily with *any* form of understanding. They are, as we shall see, part and parcel of any organization or ordering of data that we call information.

---

1 *The Idea of History* (London: Oxford University Press 1946), p. 1, and *An Essay on Philosophical Method* (London: Oxford University Press 1933), pp. 1-2.

The consequences of these limitations are varied: As a result of them, hermeneutics cannot help but be hermetic; scientific theories of necessity are circumscribed by the boundaries of the ideas that define them; formal systems must choose between consistency and comprehensiveness; philosophical study, because it includes itself within its own proper subject matter, is forced to be reflexive in its self-enclosure. The fundamental *dynamic* shared by all forms of understanding testifies to an internal limitative keystone.

Kant's architectonic suggested the existence of this keystone, which was expressed in his theory of subjective constitution, the molding of the world by the built-in categories of human intelligibility. Whorf's study of natural languages sought to make this keystone apparent in his theory of linguistic relativity, which proposed that what we can grasp is limited by the expressive capacities of our language. Gödel, Löwenheim, Skolem, and others felt its presence in the world of deductive proof, in a variety of forms that recapitulate linguistic relativity on a formal level. Husserl and his student, Eugen Fink, seemed to recognize its reality in the self-contained nature of the phenomenological attitude, which requires a basic leap from an intuitive, "naturalistic" understanding of the world, to a conceptual *conversion* that brings with it an essentially distinct approach to self-understanding. In large- and in small-scale physics, aspects of the same limitative dynamic are visible in both relativity theory and quantum mechanics, in the form of systematic acknowledgements that physical reality is intrinsically defined as a function of the observer's framework and state.[2]

Not by any concerted act of imagination can we trespass beyond the boundaries of what for us is imaginable. — This is a tight tautology, within which we realize all the intellectual freedom that is possible for us. The internal limitations of human understanding disclose themselves in several distinct ways: In our practical dealings with the world, we are subject to *neurological* limitations and to limitations of *language* and *idea*. And in our conceptual efforts, we are constrained by *epistemological* boundaries.

The limitations that structure our practice are set in place, and yet also are revealed to us, by human neurology, by the range of concepts available to us, and in part by the structure of human natural and formal language. Our neurology, conceptual vocabulary, and linguistic resources all are *encoding systems* that provide us with the spectrum or palette of colors in terms of which the world we inhabit develops its reality. Human neurology defines what we are able to apprehend and to which we can respond; the range of our concepts and the structure of our language enable us to think and to talk — within the elastic boundaries we must ever carry with us. Beyond these limitations ingredient in our encoding *abilities*, there are certain epistemologi-

---

2 These and other examples are discussed in the author's introductory essay, "Varieties of Self-Reference," in Steven J. Bartlett and Peter Suber, eds., *Self-Reference: Reflections on Reflexivity* (Dordrecht, Holland: Martinus Nijhoff 1987).

cal boundaries, which we will touch upon in a moment, that define the limits of *possible knowledge*.

The picture of the human condition suggested by these limitative factors is one of a finite organism whose neurology is responsive to a range of possible stimuli, whose conceptual vocabulary permits a certain breadth of theoretical representation, whose natural and abstract languages allow for a scope of expression and demonstration, and whose extent of knowledge is determined by conditions and limits described by epistemology. This is a picture of a creature who inhabits a specifically human universe of meaning, one that seems to be a fragment — a larger or a smaller fragment, but a fragment nonetheless — of a more inclusive reality, from contact with which our practical and theoretical limitations eternally bar us: what has, in short, been called "noumenal reality."

Appealing though this picture may be to poetic inspiration, it is a grossly distorted one: It misconstrues the compass and the kind of internal limitation that is our subject here. This view, which situates human reality within a more comprehensive framework, *exports* and yet *presupposes* the very concepts, language, and neurology that define the human perspective. In this step of exportation, we run headlong into the invisible netting of epistemology's constraints, from which "escape" is not only impossible but, on reflection, also is unthinkable. The existence of these constraints is theoretically determined, and does not depend upon the contingent biological, conceptual, or linguistic abilities of a particular organism: In attempting to refer beyond the reality made possible by our neurology, concepts, and language, we attempt, in essence, to refer beyond the reach of our referring capacities. We seek to do the impossible — not the impossible in practice — but the impossible in principle.

The so-called "boundaries" of our understanding are very peculiar limits, unlike the boundaries that delimit a field, or the walls that enclose a box. They much more closely resemble the self-limiting and yet unbounded character of a continuum that has no "outside," such as is formed by a topologically recurved surface or volume. A close analogy is the relativistic model of the physical universe, unbounded yet finite. In such a model, no matter where one goes, no matter how far, there is no way "out." For the very notion of an "outside" is *part* of the universe of meaning whose internal limitations we may now perhaps begin to appreciate. These "limitations" are of a special, philosophical variety; here the ordinary meaning of the word has undergone a radical change.

If we cannot reasonably assert (or deny) that there is an "outside," lying "beyond the reach" of the powers of our neurologies, concepts, and languages, then does it in fact make sense to say that we are *constrained* by these alleged internal limitations? Where do these limits, which we cannot meet, touch, or see, reside? Is it merely a Procrustean stretching of language to suggest that these are "limits" at all?

SELF-REFERENCE AS A TOOL OF CONCEPTUAL ANALYSIS

Internal limitations of this kind cannot by direct assault be coerced to show themselves. Yet their presence will be made evident to us, again and again, and often with little effort, if we employ an indirect strategy.

Hints of the internal limitations of human understanding came first in the guise of paradoxes involving self-reference. For centuries, these were dismissed as sophistry, no more than interesting philosophical parlour tricks. But then they became serious when suddenly, at the turn of the century, a number of paradoxes in number theory and the theory of classes were produced through the use of reflexive strategies. Other paradoxes were soon discovered, usually through the reflexive application of certain inconsistency-engendering predicates. Within a dozen years this family of paradoxes came to include the Burali-Forti paradox (1897), Cantor's paradox (1899), Russell's paradox (1901), the Richard paradox (1905), the Zermelo-König paradox (1905), Berry's paradox (1908), Grelling's paradox (1908), followed by others.

For a time, the fact that reflexive or self-referential techniques could lead to paradox brought criticism to bear on the use of self-reference itself: If we simply shunned reflexivity, we might be spared the intellectual inconvenience of paradox.[3] But the phenomenon refused to recede off stage.

Beginning in 1931, when Kurt Gödel's famous paper on formally undecidable propositions was published, a rash of results broke out, all relating to the discovery of internal limitations of formal deductive systems. Again, the main tools used were reflexive. Numerous theorems of formal limitation were proved — by Kleene, Roser, Kalmár, Gentzen, Church, Turing, Post, Tarski, Mostowski, Löwenheim, Skolem, Henkin, Wang, Curry, Myhill, Chwistek, Uspenskij, Kreisel, and others.

The ripples from these reflexively acquired results quickly spread to the then-new field of cybernetics, and, in time, on to its most recent progeny, among them, systems theory, information theory, and artificial intelligence. The central ideas of positive and negative feedback and feedforward were developed and applied to a growing range of topics, from research concerned with self-regulating and self-correcting systems, to studies of the human brain, psychotherapeutic interventions, and biological homeostasis.[4]

---

[3] For the present, the terms 'self-reference' and 'reflexivity' are used interchangeably. As will be seen in this volume, contributors to the literature differ in their preference for one term or the other; often 'reflexivity' appears to be the more general term, but no consensus has formed.

[4] An extensive bibliography of more than 1,200 works relating to self-reference, prepared by Peter Suber, will be found in Bartlett and Suber, *op. cit.*

In retrospect, the self-referential techniques used in mathematical logic and the foundations of mathematics were employed in what seems to have been an almost intuitive fashion. The reflexive strategies they exemplified have had to wait for a metatheory to clarify the underlying unselfconscious practice. To some extent, reflexive studies and applications in philosophy, which began to flourish in the two, three, and four decades following the discovery of the formal paradoxes, were more methodologically self-aware.

In philosophy, the phenomenon of self-reference has inspired research in three main areas: in semantic theory, theory of argument, and theory of knowledge. Of these, the earliest studies of reflexivity were made in semantic theory. They sought to understand the impact of the paradoxes encountered in number theory and the theory of classes upon the capacity of propositions, both those of formal systems and those in non-formalized discourse, to assert truth without self-referential inconsistency. Papers in Part I of this collection share this focus.

Somewhat later, a small group of philosophers began to cultivate an explicit interest in the use of self-reference in philosophical argument. Although individual examples of reflexive argumentation have peppered the history of philosophy, it was not until the middle of the twentieth century that efforts were made to construct a theory of self-referential argumentation, by examining a specific variety of self-reference that has come to be called "pragmatical" or "performative." Papers in Part II study this topic.

A third area of philosophical interest inspired by reflexivity evolved from the Kantian and Husserlian attempts to identify the transcendental preconditions of objective knowledge. Here, the internal limitations of human self-understanding become especially evident, in the human effort to acquire knowledge about the limits of what human subjects can know. The variety of self-reference relevant to this task has come to be called "metalogical." Studies of metalogical self-reference describe the general and necessary conditions that underlie our abilities, in principle, to refer at all, no matter what the object may be to which reference is made. Papers in Part III relate to with this area of study.

A fourth area of interest in reflexivity, closely allied to the three just mentioned, has recently developed in artificial intelligence. Here, philosophy has continued a long tradition as mother to a succession of disciplines: From a historical point of view, investigations of computational reflexivity in artificial intelligence grew out of studies of self-reference in formal systems — undertaken by researchers in mathematical logic, foundations of mathematics, and semantic theory, disciplines all of philosophical origin.

Research in artificial intelligence attempts to simulate certain human abilities (at present, more easily formulated, elementary abilities), which frequently are reflexive in nature, in a context in which computational capacities can surpass in speed and complexity those of their human creators. Studies of reflexivity in artificial intelligence have sought, for example, to develop computer programs enabling a non-human electronic system to coordinate facts, establish connections among them, and on

this basis to generate logically necessary, or plausible, inferences about the world. There has been a growing realization among researchers that such computational languages indeed must themselves constitute reflexive representations of reality, since the representations they make possible form part of the reality to be understood; and so we again encounter our topic, in another form. A programming language capable of general, reflexive intelligence immediately poses the need for self-referential abilities, to allow a machine to reflect on the usage of the language by the very machine whose functioning is defined by it. Papers in Part IV are variously concerned with problems in this area, and describe several ground-breaking proposals.

Finally, papers in Part V illustrate applications of a number of techniques of self-referential argumentation.

Studies of the various forms of reflexivity — semantical, pragmatical, metalogical, and computational — have contributed, as papers in this volume will make clear, to the task of making the intangible limits of understanding more clearly manifest to us who are constrained by them.

THE CRITICAL AND CONSTRUCTIVE USES OF SELF-REFERENCE

In keeping with philosophy's first commission, to provide a critical propaedeutic to self-assured clarity, philosophical applications of self-reference, on which I will now focus, have tended for the most part to be critical, negative, or corrective, seeking to identify and eliminate internally inconsistent dogmas. Some applications, however, have sought to use self-referential approaches to establish results non-destructively.

PRAGMATICAL SELF-REFERENCE

> *[M]y argument does not, at least in any obvious way, miss the point of anyone who might contend that philosophical statements can be true or false independently of the arguments used to establish or disestablish them. It acquires its force precisely from the force of this contention; for the contention can only take the form of an argument, and this very argument will at once serve as a further illustration of the thesis I have been advocating.*
>
> — Henry W. Johnstone, Jr.[5]

Pragmatical self-reference directs attention to the factual commitments involved in making an assertion. For example, the assertion, "knowledge is impossible in this

---

5 Henry W. Johnstone, Jr., *Philosophy and Argument* (University Park, Pennsylvania: Pennsylvania State University Press 1959), p. 81.

world of flux," is pragmatically self-referentially inconsistent: Provided that the assertion is in fact linked to an underlying commitment that places it in the category of knowledge-claims, the assertion is self-falsifying. The challenging task of the pragmatical self-referential analyst is to reveal the existence of the factual commitments that underlie everyday and philosophical discourse. His results stand or fall depending on the convincingness of his factually-focused demonstration.

THE CRITICAL USE OF PRAGMATICAL SELF-REFERENCE

Every philosophical system is subject to the obligation of accounting for its own possibility; it must at least be able to give such an account in its own terms. Less radically expressed, there must be no incompatibility between the doctrinal content of a philosophical theory, that which is maintained and asserted in it, on the one hand, and, on the other, the mere fact of the formulation of the theory in question. An incompatibility of such a kind would provide the basis for a decisive argument against the theory beset by that incompatibility.

— Aron Gurwitsch[6]

The strange thing is that philosophers should have been able to hold sincerely, as part of their philosophical creed, propositions inconsistent with what they themselves knew to be true; and yet, as far as I can make out, this has really happened.

— G. E. Moore[7]

Pragmatical applications of self-reference have attempted to show that such claims as these are self-falsifying:

- Pleasure is the chief good, since any good thing is made more desirable by the addition of pleasure.[8]
- The materialist can explain the causes of our ideas in terms of external bodies.[9]

---

6 Aron Gurwitch, "An Apparent Paradox in Leibnizianism," *Social Research*, Vol. 33, No. 1, 1966, p. 47.

7 G. E. Moore, "A Defense of Common Sense," *Classics of Analytic Philosophy*, ed. R. R. Ammerman (New York: McGraw-Hill 1965), pp. 53-54.

8 Argument from Eudoxus; see treatment by H. W. Johnstone, Jr., *op. cit.*, pp. 64ff.

9 H. W. Johnstone, Jr., *op. cit.*, pp. 67ff.

- Every event must have a cause.[10]
- All knowledge, including this, is a product of an organism's adjustment to its environment.[11]
- All meaningful statements are verifiable.[12]
- Science is incapable of objectivity.[13]
- The shift from one theory to another involves an incommensurable change in the meanings of the terms used, so that there cannot be any statements invariant across theories.[14]
- No hypothesis can be immune to revision.[15]
- No hypothesis can be irrevocably falsified.[16]
- All our statements lack significance.[17]

To this short list could be added many other examples, for numerous philosophical positions have been indicted for falling victims to the pragmatical variety of self-referential inconsistency. Among those that have been attacked in this way are the coherence theory of truth,[18] pragmatism,[19] scepticism,[20] intuitionism,[21] behaviorism,[22]

---

10 Argument from Hume; see discussion in Johnstone, *op. cit.*, p. 95.

11 W. M. Urban, *Beyond Realism and Idealism* (London: Allen and Unwin 1949), p. 236, and the discussion in Johnstone, *op. cit.*, pp. 69ff.

12 Richard Rorty, "The Limits of Reductionism," in J. Lieb, ed., *Experience, Existence, and the Good* (Carbondale: Southern Illinois University Press 1961), pp. 100-116; cf. esp. pp. 104-107.

13 Carl R. Kordig, "Objectivity, Scientific Change, and Self-Reference," in *Boston Studies in the Philosophy of Science*, Vol. 8 (Dordrecht, Holland: D. Reidel 1970), pp. 519-523.

14 *Ibid.*

15 *Ibid.*

16 *Ibid.*

17 John Passmore, *Philosophical Reasoning* (London: Duckworth 1961), p. 69.

18 E. G. Spaulding, *The New Rationalism* (New York: Holt 1918), pp. 350-351.

19 Josiah Royce, "The Eternal and the Practical," *Philosophical Review*, Vol. 13, 1904, pp. 128-129.

20 W. M. Urban, *The Intelligible World* (London: Allen and Unwin 1929), pp. 45-46, and John Passmore, *op. cit.*, pp. 72ff.

21 W. E. Hocking, *Types of Philosophy* (New York: Charles Scribner's Sons 1939), p. 201.

22 A. O. Lovejoy, "The Paradox of the Thinking Behaviorist," *Philosophical Review*, Vol. 31, 1922, pp. 142-147.

determinism,[23] subjectivism,[24] views that oppose idealism,[25] and views that oppose utilitarianism.[26]

THE CONSTRUCTIVE USE OF PRAGMATICAL SELF-REFERENCE

[V]alid constructive arguments in philosophy must in fact be circular. . . . All valid constructive philosophical arguments involve this element of feedback.

— Henry W. Johnstone, Jr.[27]

In contrast to the preceding critical arguments that utilize pragmatical self-reference to undercut a disagreeable thesis, a few philosophers have tried to use the approach constructively.

(We should remark that this distinction, between critical and constructive arguments, admittedly is often difficult to draw clearly, especially in the present context: A pragmatically critical argument establishing that *P* is self-falsifying leads to the conclusion *not-P*; yet, if *not-P* is thought to be a philosophically significant result, the argument's proponent naturally believes his argument is constructive. Among arguments and their proponents, the constructiveness of their conclusions can be stretched across a broad spectrum. At the dim end of lesser interest one might place, for instance, the critical argument against the assertion, "All our statements lack significance." The self-referential argument that establishes the negation of this assertion resists being thought of as especially interesting or constructive. Certainly it tells us something of which few are ignorant.)

In general, constructive self-referential argumentation attempts to demonstrate a positive thesis, rather than to undermine an erroneous view maintained by someone else: Admittedly, when it comes to showing that others are wrong, the judo-like strategy of utilizing feedback in argumentation is especially well-suited, as a reader

---

23 J. R. Lucas, *Freedom of the Will* (Oxford: Clarendon Press 1970) and J. M. Boyle, Jr., G. Grisez, and O. Tollefsen, *Free Choice: A Self-Referential Argument* (Notre Dame, Indiana: University of Notre Dame Press 1976).

24 An argument originally advanced by Protagoras: see treatment in John Passmore, *op. cit.*, pp. 64ff.

25 Josiah Royce, *Lectures on Modern Idealism* (New Haven, Connecticut: Yale University Press 1919), pp. 237-240.

26 Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation* (Oxford: Clarendon Press 1876), Chap. 1, sections 13-14.

27 H. W. Johnstone, Jr., *op. cit.*, pp. 76, 78.

new to the field intuitively may suspect. But some constructive arguments have, nevertheless, been formulated using the tools of pragmatical self-reference. A few we might mention here are:

- Moore's defense of common sense, using its appeal;[28]
- The argument that there are invariant conditions of discourse;[29]
- Arguments seeking to demonstrate the ontological commitments of discourse, and the related argument claiming that all objects of which we are conscious are, in diverse senses, real;[30]
- The self-confirming evidence that a sound is audible, *is* that we hear it;[31]
- The defense of "orientational pluralism" in philosophy: According to this view, philosophical positions represent relativistic frames of reference. For them, there is no *unique* solution to philosophical problems.[32]

To these examples may be added the larger group of arguments that progress from a self-referential refutation of an opposing thesis to the affirmation of its philosophically significant negation. Among these are found the positions mentioned earlier that defend: the objectivity of science, free choice, utilitarianism, idealism, the thesis that verifiability is not a property belonging to all meaningful statements, etc.

METALOGICAL SELF-REFERENCE

[W]e are brought to the conclusion that we can never transcend the limits of possible experience.

— Immanuel Kant[33]

---

28 J. Passmore, *op. cit.*, pp. 78ff.

29 J. Passmore, *op. cit.*, pp. 69ff; Paul Lorenzen, *Normative Logic and Ethics* (Mannheim/Zürich: Bibliographisches Institut 1969), p. 14, and (in connection with operative logic) p. 89. See also Lorenzen's *Einführung in die operative Logik und Mathematik* (Berlin: Springer Verlag 1969).

30 W. V. O. Quine, *Ontological Relativity* (New York: Columbia University Press 1969), and Alexius Meinong, "The Theory of Objects," in Roderick M. Chisholm, ed., *Realism and the Background of Phenomenology* (Glencoe, Illinois: The Free Press 1960), pp. 76-117.

31 Mentioned by J. S. Mill, "Proof of the Principle of Utility," in *Utilitarianism* (Indianapolis, Indiana: Bobbs-Merrill 1971; originally published 1863), Chapter IV; also see discussion in H. W. Johnstone, Jr., *op. cit.*, pp. 77ff.

32 Nicholas Rescher, "Philosophical Disagreement: An Essay towards Orientational Pluralism in Metaphilosophy," *Review of Metaphysics*, Vol. 32, No. 2, 1979, pp. 217-251.

33 *Critique of Pure Reason*, Preface to the Second Edition, B xix.

Unlike strategies of argumentation using pragmatical self-reference, metalogical approaches direct attention to the commitments that are necessarily involved if a claim or concept *in principle* is to be capable of referring to those objects to which reference is presupposed. Philosophical argument that relies upon principles of pragmatical reflexivity is usually and appropriately considered to be *ad hominem* in nature. Reflexive argumentation developed on a metalogical basis, on the other hand, has an unmistakable transcendental orientation.

Universally, for a claim to function as such it must refer to certain objects, about which assertion is made. Metalogical reflexivity comes to be of interest in connection either, from a critical point of view, with claims that conflict with their own referential preconditions, or, from a constructive point of view, with claims that compel assent, since they cannot be denied without producing such a conflict.

ITS CRITICAL USE

[O]ne must avoid the error of assuming that the sense behind familiar notions is obvious.

— D. C. Ipsen[34]

It constitutes a great advance in our critical attitude. . . to realize that a great many of the questions that we uncritically ask are without meaning. . . . [O]ne is making a significant statement about his subject in stating that a certain question is meaningless.

— P. W. Bridgman[35]

Metalogical applications of self-reference have attempted to identify a wide range of self-undermining concepts and claims. Among these are:

- Descartes' methodologically sceptical hypothesis of an evil genius, capable of shaking all confidence in our abilities to ascertain the truth about reality;[36]

---

34 D. C. Ipsen, *Units, Dimensions, and Dimensionless Numbers* (New York: McGraw-Hill 1960), p. v.

35 P. W. Bridgman, *The Logic of Modern Physics* (New York: Macmillan 1961; first printed 1927), pp. 28-29.

36 O. K. Bouwsma, "Descartes' Evil Genius," in Alexander Sesonske and Noel Fleming, eds., *Meta-meditations: Studies in Descartes* (Belmont, California: Wadsworth Publishing Co. 1965), pp. 26-36; and Steven J. Bartlett, "Hoisted by Their Own Petards: Philosophical Positions that Self-Destruct," *Argumentation*, Vol. 2, 1988, pp. 221-232.

- Kant's distinction between objects spatially structured by the human mind and "objects themselves," to which the human concept of space does not apply;[37]

- The hidden variable interpretation of quantum mechanics, which expressed a bias in favor of realism and physical determinism on the level of small-particle interactions;[38]

- Philosophical scepticism as treated by P. F. Strawson;[39]

- The argument (which ironically depended on a pragmatically reflexive strategy) attempting to show that the *rejection* of free choice is self-falsifying, or else pointless;[40]

- The view claiming that solutions to mathematical or other problems are "discovered"; they are not "invented";

- The opposing view, claiming that solutions to mathematical or other problems are "invented"; they are not "discovered";[41]

- The doctrine that there exists (or does not exist) a "metaphysical self";[42]

- The belief that a phenomenological description of an experience tells us what was "already present" in the experience pre-reflectively and implicitly;[43]

---

37 See Bartlett, *ibid.*

38 Cf. Steven J. Bartlett, "Self-Reference, Phenomenology, and Philosophy of Science," *Methodology and Science*, Vol. 13, No. 3, 1980, section VII.

39 See Strawson's argument against scepticism in *Individuals*, which can be interpreted as an attempt to prove that the sceptic's position is metalogically self-undermining.

40 This argument was advanced in J. M. Boyle, Jr., G. Grisez, and O. Tollefsen in *Free Choice: A Self-Referential Argument* (Notre Dame, Indiana: University of Notre Dame Press 1976). Although a hard-working attempt to show that freedom of choice may be rejected only on pain of pragmatical self-referential inconsistency or pointlessless, the argument itself is metalogically self-undermining. See Bartlett, review of Boyle-Grisez-Tollefsen's book, in *Review of Metaphysics*, Vol. 32, No. 4, 1979, pp. 738-740.

41 On this hypothesis and the preceding one, see Bartlett, "A Metatheoretical Basis for Interpretations of Problem-Solving Behavior," *Methodology and Science*, Vol. 11, No. 2, 1978, pp. 59-85: esp. pp. 70-72, 79-82.

42 Steven J. Bartlett, "The Use of Protocol Analysis in Philosophy," *Metaphilosophy*, Vol. 9, Nos. 3 and 4, 1978, pp. 324-336.

43 Steven J. Bartlett, "Phenomenology of the Implicit," *Dialectica*, Vol. 29, Nos. 2-3, 1975, section III,

- The Newtonian concepts of absolute time and space;[44]
- The realist view that accords a separable existence to past or future events, independently of the present;[45]
- The framework-independent concept of absolute truth;[46]
- The doctrine claiming that every event is the effect of a prior cause, and the related doctrine claiming that in a cause-effect sequence, the occurrence of the cause was indispensable to the occurrence of the effect;[47]
- The interrelated beliefs that there is a common "pole," called "the ego," shared by all of the investigator's experiences; that consciousness is a universal attribute of experience; that consciousness is a kind of "container" of experiences, beyond which meaningful claims may be made;[48]
- the doctrine that mental events are in many instances the results of prior acts (a belief inspired by the causal dogma mentioned earlier);[49]
- The belief that reflection does (or does not) perturb the structure or nature of pre-reflective experience;[50]

THE CONSTRUCTIVE USE OF METALOGICAL SELF-REFERENCE

[E]very true proposition attributing a predicate to a subject is purely analytic, since the subject is its own nature.

— Bertrand Russell[51]

---

and "Fenomenologia Tego, Co Implikowane," *Roczniki Filozoficzne*, Vol. XXII, No. 1, 1974, pp. 73-89.

44 Steven J. Bartlett, *A Relativistic Theory of Phenomenological Constitution: A Self-Referential, Transcendental Approach to Conceptual Pathology* (Université de Paris, 1970; *Diss. Abs. Internatl.* No. 7905583), Chapter 2.1.

45 *Ibid.*

46 *Ibid.*, Chapter 2.4.

47 *Ibid.*, Chapter 2.5.

48 *Ibid.*, Chapter 2.6.

49 *Ibid.*, Chapter 2.7.

50 *Ibid.*, Chapter 2.7.

51 Bertrand Russell, "The Monistic Theory of Truth," in *Philosophical Essays* (New York and London: Longmans, Green, and Co. 1910), p. 167.

The constructive use of metalogical self-reference depends upon a special property of claims of a certain kind: This is the property possessed by a claim that is such that its denial leads exactly to the variety of self-referential inconsistency that is in view here, i.e., self-referential inconsistency that *precludes* that the intended reference of the claim is possible at all.

Claims of this kind are *self-validating*, since, if they are rejected, they succumb to self-referential inconsistency of such magnitude that their capacity to be meaningful is short-circuited. As in the case of pragmatically reflexive arguments, there is an interplay between the critical and the constructive ends to which metalogically reflexive arguments may be put. The relation between criticism and construction is similarly bridged here by a conditional: If it can be shown that a claim is metalogically self-undermining, then the *rejection* of that claim will compel assent. It is important to notice that the rejection of a claim does not entail the positive endorsement of its negation. For example, the rejection of "there exists a metaphysical self" does not commit us to "one does not exist." — Both claims employ a framework-transcending concept that stands in conflict with its framework-relative basis.

In a similar way, the rejection of a self-validating claim is self-undermining.

Among positions and arguments that have sought their own validations in ways closely akin to a metalogically reflexive strategy, these could be listed:

- Kant's transcendental deduction;
- Collingwood's absolute presuppositions of systematic thought, which are presupposed by any cognition, and make knowledge possible;
- Husserl's conception of transcendental phenomenology, the analysis of which reflexively discloses the necessary foundation for its own possibility;
- Strawson's attempt to deduce, in a quasi-transcendental manner, the necessary and basic structure of a conceptual system that makes objective knowledge possible;
- Gaston Isaye's transcendental method of retortion, which seeks to identify the conditions of the possibility of thought by means of a strategy to show that every possible denial of a self-justifying assertion leads to a self-referentially inconsistent position;[52]

52 See Gaston Isaye, "La Justification critique par rétorsion," *Revue Philosophique de Louvain*, Vol. 52, 1954, pp. 205-233; Otto Muck, *The Transcendental Method*, trans. by William D. Seidensticker (New York: Herder and Herder 1968), pp. 163-180; and Martin X. Moleski, "Retortion: The Method of Gaston Isaye," *International Philosophical Quarterly*, Vol. 17, No. 1, pp. 59-83. (Moleski refers Isaye's reflexive approach to the category of pragmatical self-reference, which is clearly a mistake, given Isaye's self-conscious focus on necessary, non-contingent truths that are conditions of the possibility of human thought.)

> The following pair of mutually reinforcing positions: The author's reflexive argument that metalogical referential consistency is a necessary condition of meaning, on the one hand, and his relativistic theory of the constitution of experience, on the other. Together, these approaches show that a wide range of everyday and technical concepts is metalogically self-undermining, underscoring the need for a vocabulary of radically different but referentially self-consistent concepts.[53]

• • •

The experiential and conceptual space we inhabit has a strange structure and a surprising logic that we are only beginning to accept and appreciate. The history of physics serves almost as a parable, for it tells a larger story through having come full circle, from primitive anthropocentrism, to the displacement of man as the center around which all things revolve, to an observer-based awareness of relativity. In this recent return to framework-relativity, self-reference has played an increasingly important role across many disciplines, as a tool of discovery and analysis, and as a phenomenon worthy of study in its own right, whether in cognitive science, artificial intelligence, general systems theory, the foundations of mathematics, epistemology, or other fields.

Both on the level of our factual dealings with the contingent world, and on the abstract level of theoretical necessities, the study of reflexivity has progressed within the last hundred years from a parlour curiosity to an indispensable and perhaps the most basic tool enabling us to gain an understanding both of ourselves as well as of systems whose dynamic, like our own, appears to be fundamentally self-regulating, and self-limiting.

• • •

The essays contained in this volume consist of thirty-two papers on reflexivity, papers that form the classical basis for current research in this steadily growing area of study. They cover more than half of a century of work that was inspired, directly or indirectly, by the intellectual misgivings and confusion that followed the discovery of the semantical and set-theoretical paradoxes. These papers were originally published in numerous journals and volumes of conference proceedings, and have been brought together here for the first time, where they are printed in facsimile.

---

53 Cf., *inter alia*, Steven J. Bartlett, "Referential Consistency as a Criterion of Meaning," *Synthese*, Vol. 52, 1982, pp. 267-282, reprinted in this volume; and *A Relativistic Theory of Phenomenological Constitution: A Self-Referential, Transcendental Approach to Conceptual Pathology* (Université de Paris, 1970; *Diss. Abs. Internatl.* No. 7905583).

The individual essays were chosen for inclusion in this collection with three criteria in view: that each, when read in conjunction with others, should throw light on the evolution of thought about reflexivity; that each paper should, as we look back on the past sixty years of research, be recognizable as a basic contribution to current research; and that each article should point the interested reader on to other key contributions in the literature.

These essays, divided into families according to the varieties of reflexivity they examine, are of fundamental importance to an understanding and appreciation of the many-faceted, pervasive phenomenon of reflexivity.

PART I

SEMANTICAL SELF-REFERENCE

PARADOX

Some self-contradictory statements are amusing; others are profoundly puzzling. A few paradoxes have called for major reconstructions of the foundations of logic and mathematics

by W. V. Quine

Frederic, the young protagonist of *The Pirates of Penzance,* has reached the age of 21 after passing only five birthdays. Several circumstances conspire to make this possible. Age is reckoned in elapsed time, whereas a birthday has to match the date of birth; and February 29 comes less frequently than once a year.

Granted that Frederic's situation is possible, wherein is it paradoxical? Merely in its initial air of absurdity. The likelihood that a man will be more than *n* years old on his *n*th birthday is as little as one to 1,460, or slightly better if we allow for seasonal trends; and this likelihood is so slight that we easily forget its existence.

May we say in general, then, that a paradox is just any conclusion that at first sounds absurd but that has an argument to sustain it? In the end I think this account stands up pretty well. But it leaves much unsaid. The argument that sustains a paradox may expose the absurdity of a buried premise or of some preconception previously reckoned as central to physical theory, to mathematics or to the thinking process. Catastrophe may lurk, therefore, in the most innocent-seeming paradox. More than once in history the discovery of paradox has been the occasion for major reconstruction at the foundations of thought. For some decades, indeed, studies of the foundation of mathematics have been confounded and greatly stimulated by confrontation with two paradoxes, one propounded by Bertrand Russell in 1901 and the other by Kurt Gödel in 1931.

As a first step onto this dangerous ground, let us consider another paradox: that of the village barber. This is not Russell's great paradox of 1901, to which we shall come, but a lesser one that Russell attributed to an unnamed source in 1918. In a certain village there is a man, so the paradox runs, who is a barber; this barber shaves all and only those men in the village who do not shave themselves. Query: Does the barber shave himself?

Any man in this village is shaved by the barber if and only if he is not shaved by himself. Therefore in particular the barber shaves himself if and only if he does not. We are in trouble if we say the barber shaves himself and we are in trouble if we say he does not.

Now compare the two paradoxes. Frederic's situation seemed absurd at first, but a simple argument sufficed to make us acquiesce in it for good. In the case of the barber, on the other hand,

the conclusion is too absurd to acquiesce in at any time.

What are we to say to the argument that goes to prove this unacceptable conclusion? Happily it rests on assumptions. We are asked to swallow a story about a village and a man in it who shaves all and only those men in the village who do not shave themselves. This is the source of our trouble; grant this and we end up saying, absurdly, that the barber shaves himself if and only if he does not. The proper conclusion to draw is just that there is no such barber. We are confronted with nothing more mysterious than what logicians have been referring to for a couple of thousand years as a *reductio ad absurdum*. We disprove the barber by assuming him and deducing the absurdity that he shaves himself if and only if he does not. The paradox is simply a proof that no village can contain a man who shaves all and only those men in it who do not shave themselves. This sweeping denial at first sounds absurd; why should there not be such a man in a village? But the argument shows why not, and so we acquiesce in the sweeping denial just as we acquiesced in the possibility, absurd on first exposure, of Frederic's being so much more than five years old on his fifth birthday.

Both paradoxes are alike, after all, in sustaining prima facie absurdities by conclusive argument. What is strange but true in the one paradox is that one can be $4n$ years old on one's nth birthday; what is strange but true in the other paradox is that no village can contain a man who shaves all and only those men in the village who do not shave themselves.

Still, I would not limit the word "paradox" to cases where what is purportedly established is true. I shall call these, more particularly, veridical, or truth-telling, paradoxes. For the name of paradox is suited equally to falsidical ones. (This word is not so barbarous as it sounds; *falsidicus* occurs twice in Plautus and twice in earlier writers.)

The Frederic paradox is a veridical one if we take its proposition not as something about Frederic but as the abstract truth that a man can be $4n$ years old on his nth birthday. Similarly, the barber paradox is a veridical one if we take its proposition as being that no village contains such a barber. A falsidical paradox, on the other hand, is one whose proposition not only seems at first absurd but also is false, there being a fallacy in the purported proof. Typical falsidical paradoxes are the comic misproofs that $2=1$. Most of us have heard one or another such. Here is the version offered by the 19th-century English mathematician Augustus De Morgan: Let $x=1$. Then $x^2=x$. So $x^2-1=x-1$. Dividing both sides by $x-1$, we conclude that $x+1=1$; that is, since $x=1$, $2=1$. The fallacy comes in the division by $x-1$, which is 0.

Instead of "falsidical paradox" could I say simply "fallacy"? Not quite. Fallacies can lead to true conclusions as well as false ones, and to unsurprising conclusions as well as surprising ones. In a falsidical paradox there is always a fallacy in the argument, but the proposition purportedly established has furthermore to seem absurd and to be indeed false.

Some of the ancient paradoxes of Zeno belong under the head of falsidical paradoxes. Take the one about Achilles and the tortoise. Generalized beyond these two fictitious characters, what the paradox purports to establish is the absurd proposition that so long as a runner keeps running, however slowly, another runner can never overtake him. The argument is that each time the pursuer reaches a spot where the pursued has been, the pursued has moved a bit beyond. When we try to make this argument more explicit, the fallacy that emerges is the mistaken notion that any infinite succession of intervals of time has to add up to all eternity. Actually when an in-

finite succession of intervals of time is so chosen that the succeeding intervals become shorter and shorter, the whole succession may take either a finite or an infinite time. It is a question of a convergent series.

Grelling's Paradox

The realm of paradox is not clearly exhausted even by the veridical and falsidical paradoxes together. The most startling of all paradoxes are not clearly assignable to either of these domains. Consider the paradox, devised by the German mathematician Kurt Grelling in 1908, concerning the heterological, or nonself-descriptive, adjectives.

To explain this paradox requires first a definition of the autological, or self-descriptive, adjective. The adjective "short" is short; the adjective "English" is English; the adjective "adjectival" is adjectival; the adjective "polysyllabic" is polysyllabic. Each of these adjectives is, in Grelling's terminology, autological: each is true of itself. Other adjectives are heterological; thus "long," which is not a long adjective; "German," which is not a German adjective; "monosyllabic," which is not a monosyllabic one.

Grelling's paradox arises from the query: Is the adjective "heterological" an autological or a heterological one? We are as badly off here as we were with the barber. If we decide that "heterological" is autological, then the adjective is true of itself. But that makes it heterological rather than autological, since whatever the adjective "heterological" is true of is heterological. If we therefore decide that the adjective "heterological" is heterological, then it is true of itself, and that makes it autological.

Our recourse in a comparable quandary over the village barber was to declare a *reductio ad absurdum* and conclude that there was no such barber. Here, however, there is no interim premise to disavow. We merely defined the adjective "heterological" and asked if it was heterological. In fact, we can get the paradox just as well without the adjective and its definition. "Heterological" was defined as meaning "not true of self"; we can therefore ask if the adjectival phrase "not true of self" is true of itself. We find that it is if and only if it is not, hence that it is and it is not; and so we have our paradox.

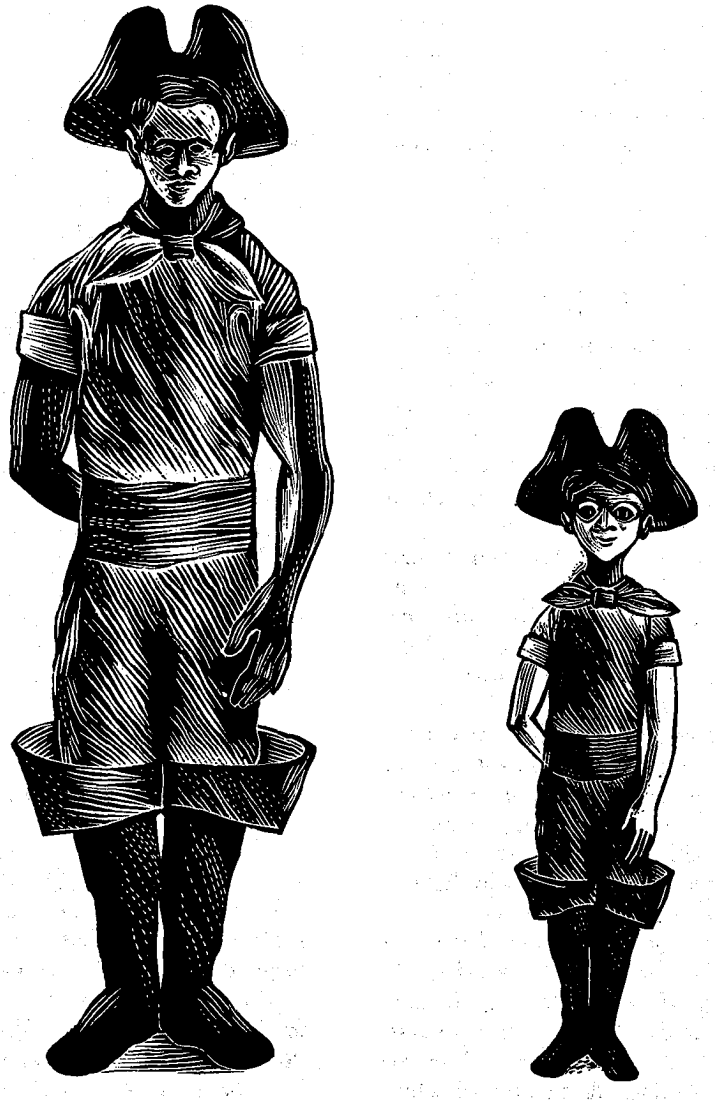
Thus viewed, Grelling's paradox seems unequivocally falsidical. Its proposition is a self-contradictory compound proposition to the effect that our adjective is and is not true of itself. But this paradox contrasts strangely with the falsidical paradoxes of Zeno, or of $2=1$, in that we are at a loss to spot the fallacy in the argument. It may for this reason be best seen as representing a third class of paradoxes, separate from the veridical and falsidical ones.

Antinomies

The paradoxes of this class are called antinomies, and it is they that bring on the crises in thought. An antinomy produces a self-contradiction by accepted ways of reasoning. It establishes that some tacit and trusted pattern of reasoning must be made explicit and henceforward be avoided or revised.

Take Grelling's paradox, in the form in which it shows the adjective phrase "not true of self" to be both true and false of itself. What tacit principles of reasoning does the argument depend on? Notably this one: the adjective "red" is true of a thing if and only if the thing is red; the adjective "big" is true of a thing if and only if the thing is big; the adjective "not true of self" is true of a thing if and only if the thing is not true of itself; and so on. This last case of the principle is the case that issues directly in the paradox.

There is no denying that this principle is constantly used, tacitly, when we speak of adjectives as true of things: the



"MOST INGENIOUS PARADOX" of *The Pirates of Penzance* involves Frederic, who was born on a February 29. He is 21, but going by birthdays "only five and a little bit over."

adjective "red" is true of a thing if and only if the thing is red, and correspondingly for all adjectives. This principle simply reflects what we mean in saying that adjectives are true of things. It is a hard principle to distrust, and yet it is obviously the principle that is to blame for our antinomy. The antinomy is directly a case of this principle. Take the adjective in the principle as the adjectival phrase "not true of self" instead of the adjective "red," and take the "thing" in the principle, of which the adjective is to be true, as that adjective over again; thereupon the principle says outright that "not true of self" is true of

itself if and only if it is not true of itself. So the principle must be abandoned or at least somehow restricted.

Yet so faithfully does the principle reflect what we mean in calling adjectives true of things that we cannot abandon it without abjuring the very expression "true of" as pernicious nonsense. We could still go on using the adjectives themselves that had been said to be true of things; we could go on attributing them to things as usual; what we would be cutting out in "true of" is merely a special locution for talking about the attribution of the adjectives to the things.

This special locution, however, has its conveniences, and it would be missed. In fact, we do not have to do without it altogether. After all, to speak of adjectives as true or not true of things makes trouble only in a special case, involving one special adjective, namely the phrase "not true of self," in attribution to one special thing, namely that same phrase over again. If we forswear the use of the locution "true of" in connection with this particular phrase in relation to itself as object, we thereby silence our antinomy and may go on blithely using the locution "true of" in other cases as always, pending the discovery of further antinomies.

Actually related antinomies are still forthcoming. To inactivate the lot we have to cut a little deeper than our one case; we have to forswear the use of "true of" not only in connection with "not true of self" but also in connection with various other phrases relating to truth; and in such connections we have to forswear the use not only of "true of" but also of various other truth locutions. First let us look at some of the antinomies that would otherwise threaten.

The Paradox of Epimenides

There is the ancient paradox of Epimenides the Cretan, who said that all Cretans were liars. If he spoke the truth,

BARBER PARADOX assumes that in a certain village there is a barber who shaves all and only those men who do not shave themselves. The question is whether this barber shaves himself. The paradox is that he does shave himself only if he does not.

he was a liar. It seems that this paradox may have reached the ears of St. Paul and that he missed the point of it. He warned, in his epistle to Titus: "One of themselves, even a prophet of their own, said, The Cretans are always liars."

Actually the paradox of Epimenides is untidy; there are loopholes. Perhaps some Cretans were liars, notably Epimenides, and others were not; perhaps Epimenides was a liar who occasionally told the truth; either way it turns out that the contradiction vanishes. Something of paradox can be salvaged with a little tinkering; but we do better to switch to a different and simpler rendering, also ancient, of the same idea. This is the *pseudomenon*, which runs simply: "I am lying." We can even drop the indirectness of a personal reference and speak directly of the sentence: "This sentence is false." Here we seem to have the irreducible essence of antinomy: a sentence that is true if and only if it is false.

In an effort to clear up this antinomy it has been protested that the phrase "This sentence," so used, refers to nothing. This is claimed on the ground that you cannot get rid of the phrase by supplying a sentence that is referred to. For what sentence does the phrase refer to? The sentence "This sentence is false." If, accordingly, we supplant the phrase "This sentence" by a quotation of the sentence referred to, we get: "'This sentence is false' is false." But the whole outside sentence here attributes falsity no longer to itself but merely to something other than itself, thereby engendering no paradox.

If, however, in our perversity we are still bent on constructing a sentence that does attribute falsity unequivocally to itself, we can do so thus: "'Yields a falsehood when appended to its own quotation' yields a falsehood when appended to its own quotation." This sentence specifies a string of nine words and says of this string that if you put it down twice, with quotation marks around the first of the two occurrences, the result is false. But that result is the very sentence that is doing the telling. The sentence is true if and only if it is false, and we have our antinomy.

This is a genuine antinomy, on a par with the one about "heterological," or "false of self," or "not true of self," being true of itself. But whereas that earlier one turned on "true of," through the construct "not true of self," this new one turns merely on "true," through the construct "falsehood," or "statement not true." We can avoid both antinomies, and others related to them, by ceasing to use "true of" and "true" and their equivalents and derivatives, or at any rate ceasing to apply such truth locutions to adjectives or sentences that themselves contain such truth locutions.

This restriction can be relaxed somewhat by admitting a hierarchy of truth locutions, as suggested by the work of Bertrand Russell and the Polish mathematician Alfred Tarski, who is now at the University of California. The expressions "true," "true of," "false" and related ones can be used with numerical subscripts "0," "1," "2," and so on always attached or imagined; thus "true_0," "true_1," "true_2," "false_0" and so on. Then we can avoid the antinomies by taking care, when a truth locution (T) is applied to a sentence or other expression (S), that the subscript on T is higher than any subscript inside S. Violations of this restriction would be treated as meaningless, or ungrammatical, rather than as true or false sentences. For instance, we could meaningfully ask whether the adjectives "long" and "short" are true_0 of themselves; the answers are respectively no and yes. But we could not meaningfully speak of the phrase "not true_0 of self" as true_0 or false_0 of itself; we would have to ask whether it is true_1 or false_1 of itself, and this is a question that leads to no antin-

omy. Either way the question can be answered with a simple and unpenalized negative.

This point deserves to be restated: Whereas "long" and "short" are adjectives that can meaningfully be applied to themselves, falsely in the one case and truly in the other, on the other hand "true$_0$ of self" and "not true$_0$ of self" are adjectival phrases that cannot be applied to themselves meaningfully at all, truly or falsely. Therefore to the question "Is 'true$_0$ of self' true$_1$ of itself?" the answer is no; the adjectival phrase "true$_0$ of itself" is meaningless of itself rather than true$_1$ of itself.

Next let us consider, in terms of subscripts, the most perverse version of the *pseudomenon*. We have now, for meaningfulness, to insert subscripts on the two occurrences of the word "falsehood," and in ascending order, thus: " 'Yields a falsehood$_0$ when appended to its own quotation' yields a falsehood$_1$ when appended to its own quotation." Thereupon paradox vanishes. This sentence is unequivocally false. What it tells us is that a certain described form of words is false$_1$, namely the form of words: " 'Yields a falsehood$_0$ when appended to its own quotation' yields a falsehood$_0$ when appended to its own quotation." But in fact this form of words is not false$_1$; it is meaningless. So the preceding sentence, which said that this form of words was false$_1$, is false. It is false$_2$.

This may seem an extravagant way of eliminating antinomies. But it would be much more costly to drop the word "true," and related locutions, once and for all. At an intermediate cost one could merely leave off applying such

EPIMENIDES THE CRETAN made the statement that all Cretans were liars. Such a statement can be simplified to "I am lying" or "This sentence is false." One can seemingly prove of such paradoxes, called antinomies, that they are true if and only if they are false.

locutions to expressions containing such locutions. Either method is less economical than this method of subscripts. The subscripts do enable us to apply truth locutions to expressions containing such locutions, although in a manner disconcertingly at variance with custom. Each resort is desperate; each is an artificial departure from natural and established usage. Such is the way of antinomies.

A veridical paradox packs a surprise, but the surprise quickly dissipates itself as we ponder the proof. A falsidical paradox packs a surprise, but it is seen as a false alarm when we solve the underlying fallacy. An antinomy, however, packs a surprise that can be accommodated by nothing less than a repudiation of part of our conceptual heritage.

Revision of a conceptual scheme is not unprecedented. It happens in a small way with each advance in science, and it happens in a big way with the big advances, such as the Copernican revolution and the shift from Newtonian mechanics to Einstein's theory of relativity. We can hope in time even to get used to the biggest such changes and to find the new schemes natural. There was a time when the doctrine that the earth revolves around the sun was called the Copernican paradox, even by the men who accepted it. And perhaps a time will come when truth locutions without implicit subscripts, or like safeguards, will really sound as nonsensical as the antinomies show them to be.

Conversely, the falsidical paradoxes of Zeno must have been, in his day, genuine antinomies. We in our latter-day smugness point to a fallacy: the notion that an infinite succession of intervals must add up to an infinite interval. But surely this was part and parcel of the conceptual scheme of Zeno's day. Our recognition of convergent series, in which an infinite number of segments add up to a finite segment, is from Zeno's vantage point an artificiality comparable to our new subscripts on truth locutions. Perhaps these subscripts will seem as natural to our descendants of A.D. 4000, granted the tenuous hypothesis of there being any, as the convergent series does to us. One man's antinomy is another man's falsidical paradox, give or take a couple of thousand years.

I have not, by the way, exhausted the store of latter-day antinomies. Another good one is attributed by Russell to a librarian named Berry. Here the theme is numbers and syllables. Ten has a one-syllable name. Seventy-seven has a five-syllable name. The seventh power of seven hundred seventy-seven has a name that, if we were to work it out, might run to 100 syllables or so; but this number can also be specified more briefly in other terms. I have just specified it in 15 syllables. We can be sure, however, that there are no end of numbers that resist all specification, by name or description, under 19 syllables. There is only a finite stock of syllables altogether, and hence only a finite number of names or phrases of less than 19 syllables, whereas there are an infinite number of positive integers. Very well, then; of those numbers not specifiable in less than 19 syllables, there must be a least. And here is our antinomy: the least number not specifiable in less than nineteen syllables is specifiable in 18 syllables. I have just so specified it.

This antinomy belongs to the same family as the antinomies that have gone before. For the key word of this antinomy, "specifiable," is interdefinable with "true of." It is one more of the truth locutions that would take on subscripts under the Russell-Tarski plan. The least number not specifiable$_0$ in less than nineteen syllables is indeed specifiable$_0$ in 18 syllables, but it is not specifiable$_0$ in less than 19 syllables; for all I know it is not specifiable$_0$ in less than 23. This resolution of Berry's antinomy is the one

ZENO'S PARADOX of Achilles and the tortoise proposes an absurdity: that so long as the tortoise continues to move, however slowly, the fleet Achilles can never overtake him. The paradox is called falsidical, there being a fallacy in its purported proof.

that would come through automatically if we paraphrase "specifiable" in terms of "true of" and then subject "true of" to the subscript treatment.

Russell's Antinomy

Not all antinomies belong to this family. The most celebrated of all antinomies, discovered by Russell in 1901, belongs outside this family. It has to do with self-membership of classes. Some classes are members of themselves; some are not. For example, the class of all classes that have more than five members clearly has more than five classes as members; therefore the class is a member of itself. On the other hand, the class of all men is not a member of itself, not being a man. What of the class of all classes that are not members of themselves? Since its members are the nonself-members, it qualifies as a member of itself if and only if it is not. It is and it is not: antinomy's by now familiar face.

Russell's antinomy bears a conspicuous analogy to Grelling's antinomy of "not true of self," which it long antedates. But Russell's antinomy does not belong to the same family as the Epimenides antinomy and those of Berry and Grelling. By this I mean that Russell's antinomy cannot be blamed on any of the truth locutions, nor is it resolved by subjecting those locutions to subscripts. The crucial words in Russell's antinomy are "class" and "member," and neither of these is definable in terms of "true," "true of" or the like.

I said earlier that an antinomy establishes that some tacit and trusted pattern of reasoning must be made explicit and be henceforward avoided or revised. In the case of Russell's antinomy, the tacit and trusted pattern of reasoning that is found wanting is this: for any condition you can formulate, there is a class whose members are the things meeting the condition.

This principle is not easily given up.

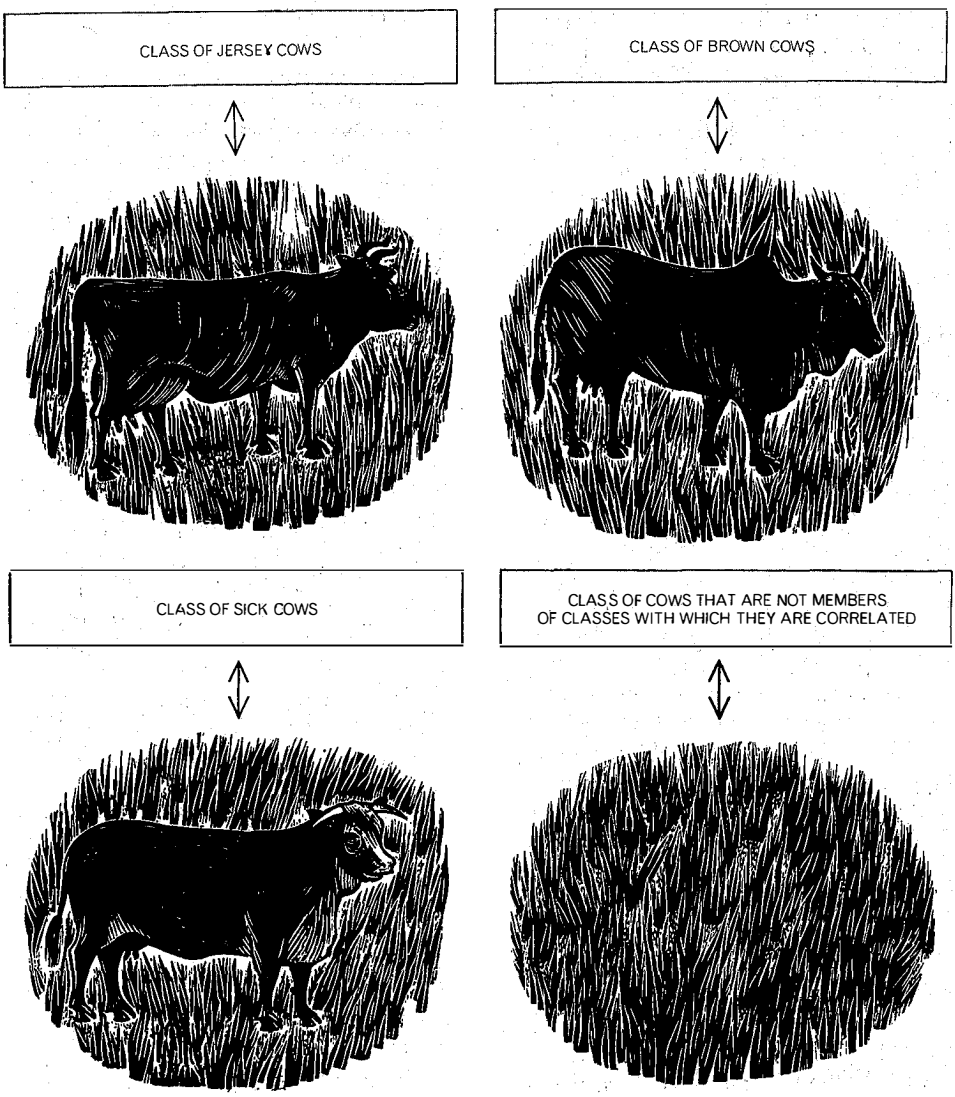
The almost invariable way of specifying a class is by stating a necessary and sufficient condition for belonging to it. When we have stated such a condition, we feel that we have "given" the class and can scarcely make sense of there not being such a class. The class may be empty, yes; but how could there not be such a class at all? What substance can be asked for it that the membership condition does not provide? Yet such exhortations avail us nothing in the face of the antinomy, which simply proves the principle untenable. It is a simple point of logic, once we look at it, that there is no class, empty or otherwise, that has as members precisely the classes that are not members of themselves. It would have to have itself as member if and only if it did not.

Russell's antinomy came as a shock to Gottlob Frege, the German mathematician who founded mathematical logic. In his *Grundgesetze der Arithmetik* Frege thought that he had secured the foundations of mathematics in the self-consistent laws of logic. He received a letter from Russell as the second volume of this work was on its way to press. "Arithmetic totters," Frege is said to have written in answer. An appendix that he added to the volume opens with the words: "A scientist can hardly encounter anything more undesirable than to have the foundation collapse just as the work is finished. I was put in this position by a letter from Bertrand Russell..."

In Russell's antinomy there is more than a hint of the paradox of the barber. The parallel is, in truth, exact. It was a simple point of logic that there was in no village a man who shaved all and only those men in the village who did not shave themselves; he would shave himself if and only if he did not. The barber paradox was a veridical paradox showing that there is no such barber. Why is Russell's antinomy then not a veridical paradox showing that there is no class whose members are all and only the nonself-members? Why does it count as an antinomy and the barber paradox not? The reason is that there has been in our habits of thought an overwhelming presumption of there being such a class but no presumption of there being such a barber. The barber paradox barely qualifies as paradox in that we are mildly surprised at being able to exclude the barber on purely logical grounds by reducing him to absurdity. Even this surprise ebbs as we review the argument; and anyway we had never positively believed in such a barber. Russell's paradox is a genuine antinomy because of the fundamental nature of the principle of class existence that it compels us to give up. When in a future century the absurdity of that principle has become a commonplace, and some substitute principle has enjoyed long enough tenure to take on somewhat the air of common sense, perhaps we can begin to see Russell's paradox as no more than a veridical paradox, showing that there is no such class as that of the nonself-members. One man's antinomy can be another man's veridical paradox, and one man's veridical paradox can be another man's platitude.

Russell's antinomy made for a more serious crisis still than did Grelling's and Berry's and the one about Epimenides. For these strike at the semantics of truth and denotation, but Russell's strikes at the mathematics of classes. Classes are appealed to in an auxiliary way in most branches of mathematics, and increasingly so as passages of mathematical reasoning are made more explicit. The basic principle of classes that is tacitly used, at virtually every turn where classes are involved at all, is precisely the class-existence principle that is discredited by Russell's antinomy.

I spoke of Grelling's antinomy and Berry's and the Epimenides as all in a family, to which Russell's antinomy does not belong. For its part, Russell's antinomy has family connections of its own.



CANTOR'S PROOF is important in set theory. He showed that there are always more classes of things of a kind than there are things of that kind. Take cows, for example, and classes of cows (***indicated here by rectangles***). If every cow is arbitrarily correlated with a class (of which it may or may not be a member), there will remain a class that is not correlated with any cow.

In fact, it is the first of an infinite series of antinomies, as follows. Russell's antinomy shows that there is no class whose members are precisely the classes that are not members of themselves. Now there is a parallel antinomy that shows there is no class whose members are precisely the classes that are not members of members of themselves. Further, there is an antinomy that shows there is no class whose members are precisely the classes that are not members of members of members of themselves. And so on ad infinitum.

All these antinomies, and other related ones, can be inactivated by limiting the guilty principle of class existence in a very simple way. The principle is that for any membership condition you can formulate there is a class whose members are solely the things meeting the condition. We get Russell's antinomy and all the others of its series by taking the condition as nonmembership in self, or nonmembership in members of self, or the like. Each time the trouble comes of taking a membership condition that itself talks in turn of membership and nonmembership. If we withhold our principle of class existence from cases where the membership condition mentions membership, Russell's antinomy and related ones are no longer forthcoming. This restriction on class existence is parallel to a restriction on the truth locutions that we contemplated for a while, before bringing in the subscripts; namely, not to apply the truth locutions to expressions containing any of the truth locutions.

Happily we can indeed withhold the principle of class existence from cases where the membership condition mentions membership, without unsettling those branches of mathematics that make only incidental use of classes. This is why it has been possible for most branches of mathematics to go on blithely using classes as auxiliary apparatus in spite of Russell's and related antinomies.

The Mathematics of Classes

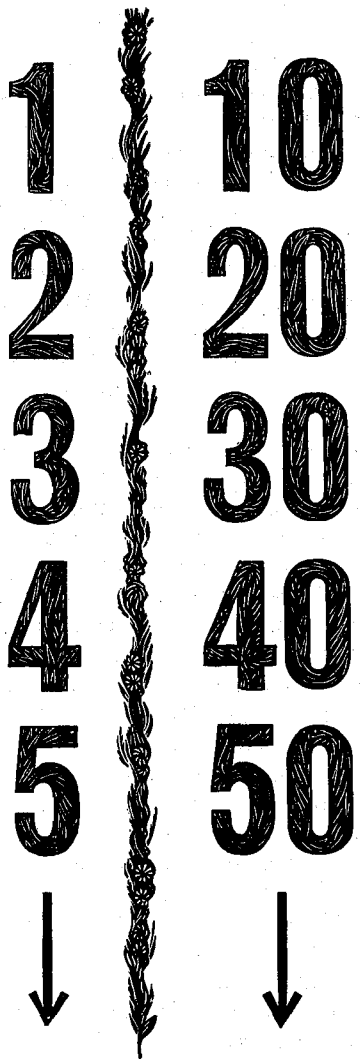
There is a particular branch of mathematics in which the central concern is with classes: general set theory. In this domain one deals expressly with classes of classes, classes of classes of classes, and so on, in ways that would be paralyzed by the restriction just now contemplated: withholding the principle of class existence from cases where the membership condition mentions membership. So one tries in general set theory to manage with milder restrictions.

General set theory is rich in paradox. Even the endless series of antinomies that I mentioned above, of which Russell's was the first, by no means exhausts this vein of paradox. General set theory is primarily occupied with infinity—infinite classes, infinite numbers—and so is involved in paradoxes of the infinite. A rather tame old paradox under this head is that you can exhaust the members of a whole class by correlating them with the members of a mere part of the class. For instance, you can correlate all the positive integers with the multiples of 10, thus: 1 with 10, 2 with 20, 3 with 30 and so on. Every positive integer gets disposed of; there are as many multiples of 10 as integers altogether. This is no antinomy but a veridical paradox. Among adepts in the field it even loses the air of paradox altogether, as is indeed the way of veridical paradox.

Georg Cantor, the 19th-century pioneer in general set theory and infinite arithmetic, proved that there are always more classes of things of a given kind than there are things of that kind; more classes of cows than cows. A distinct air of paradox suffuses his proof of this.

First note the definition of "more." What it means when one says there are more things of one kind than another is that every correlation of things of the one kind to things of the other fails to exhaust the things of the one kind. So what Cantor is proving is that no correlation of

cow classes to cows accommodates all the cow classes. The proof is as follows. Suppose a correlation of cow classes to cows. It can be any arbitrary correlation; a cow may or may not belong to the class correlated with it. Now consider the cows, if any, that do not belong to the classes correlated with them. These cows



POSITIVE INTEGERS can all be correlated with multiples of 10 even though the latter are only part of the class of integers.

themselves form a cow class, empty or not. And it is a cow class that is not correlated with any cow. If the class were so correlated, that cow would have to belong to the class if and only if it did not.

This argument is typical of the arguments in general set theory that would be sacrificed if we were to withhold the principle of class existence from cases where the membership condition mentions membership. The recalcitrant cow class that clinched the proof was specified by a membership condition that mentioned membership. The condition was nonmembership in the correlated cow class.

But what I am more concerned to bring out, regarding the cow-class argument, is its air of paradox. The argument makes its negative point in much the same way that the veridical barber paradox showed there to be no such barber, and in much the same way that Russell's antinomy showed there to be no class of nonself-members. So in Cantor's theorem —a theorem not only about cows and their classes but also about things of any sort and their classes—we see paradox, or something like it, seriously at work in the advancement of theory. His theorem establishes that for every class, even every infinite class, there is a larger class: the class of its subclasses.

So far, no antinomy. But now it is a short step to one. If for every class there is a larger class, what of the class of everything? Such is Cantor's antinomy. If you review the proof of Cantor's theorem in application directly to this disastrous example—speaking therefore not of cows but of everything—you will quickly see that Cantor's antinomy boils down, after all, to Russell's.

So the central problem in laying the foundations of general set theory is to inactivate Russell's antinomy and its suite. If such theorems as Cantor's are to be kept, the antinomies must be inactivated by milder restrictions than the total withholding of the principle of class

existence from cases where the membership condition mentions membership. One tempting line is a scheme of subscripts analogous to the scheme used in avoiding the antinomies of truth and denotation. Something like this line was taken by Russell himself in 1908, under the name of the theory of logical types. A very different line was proposed in the same year by the German mathematician Ernst Zermelo, and further variations have been advanced in subsequent years.

All such foundations for general set theory have as their point of departure the counsel of the antinomies; namely, that a given condition, advanced as a necessary and sufficient condition of membership in some class, may or may not really have a class corresponding to it. So the various alternative foundations for general set theory differ from one another with respect to the membership conditions to which they do and do not guarantee corresponding classes. Non-self-membership is of course a condition to which none of the theories accord corresponding classes. The same holds true for the condition of not being a member of any own member; and for the conditions that give all the further antinomies of the series that began with Russell's; and for any membership condition that would give rise to any other antinomy, if we can spot it.

But we cannot simply withhold each antinomy-producing membership condition and assume classes corresponding to the rest. The trouble is that there are membership conditions corresponding to each of which, by itself, we can innocuously assume a class, and yet these classes together can yield a contradiction. We are driven to seeking optimum consistent combinations of existence assumptions, and consequently there is a great variety of proposals for the foundations of general set theory. Each proposed scheme is unnatural, because the natural scheme is the unrestricted one that the antinomies discredit; and each has advantages, in power or simplicity or in attractive consequences in special directions, that each of its rivals lacks.

I remarked earlier that the discovery of antinomy is a crisis in the evolution of thought. In general set theory the crisis began 60 years ago and is not yet over.

Gödel's Proof

Up to now the heroes or villains of this piece have been the antinomies. Other paradoxes have paled in comparison. Other paradoxes have been less startling to us, anyway, and more readily adjusted to. Other paradoxes have not precipitated 60-year crises, at least not in our time. When any of them did in the past precipitate crises that durable (and surely the falsidical paradoxes of Zeno did), they themselves qualified as antinomies.

Let me, in closing, touch on a latter-day paradox that is by no means an antinomy but is strictly a veridical paradox, and yet is comparable to the antinomies in the pattern of its proof, in the surprisingness of the result and even in its capacity to precipitate a crisis. This is Gödel's proof of the incompletability of number theory.

What Kurt Gödel proved, in that great paper of 1931, was that no deductive system, with axioms however arbitrary, is capable of embracing among its theorems all the truths of the elementary arithmetic of positive integers unless it discredits itself by letting slip some of the falsehoods too [see "Gödel's Proof," by Ernest Nagel and James R. Newman; SCIENTIFIC AMERICAN, June, 1956]. Gödel showed how, for any given deductive system, he could construct a sentence of elementary number theory that would be true if and only if not provable in that system. Every such system is therefore either incomplete, in that it misses a relevant truth, or else bankrupt,

in that it proves a falsehood.

Gödel's proof may conveniently be related to the Epimenides paradox or the *pseudomenon* in the "yields a falsehood" version. For "falsehood" read "nontheorem," thus: " 'Yields a nontheorem when appended to its own quotation' yields a nontheorem when appended to its own quotation."

This statement no longer presents an antinomy, because it no longer says of itself that it is false. What it does say of itself is that it is not a theorem (of some deductive theory that I have not yet specified). If it is true, here is one truth that that deductive theory, whatever it is, fails to include as a theorem. If the statement is false, it is a theorem, in which event that deductive theory has a false theorem and so is discredited.

What Gödel proceeds to do, in getting his proof of the incompletability of number theory, is the following. He shows how the sort of talk that occurs in the above statement—talk of nontheoremhood and of appending things to quotations—can be mirrored systematically in arithmetical talk of integers. In this way, with much ingenuity, he gets a sentence purely in the arithmetical vocabulary of number theory that inherits that crucial property of being true if and only if not a theorem of number theory. And Gödel's trick works for any deductive system we may choose as defining "theorem of number theory."

Gödel's discovery is not an antinomy but a veridical paradox. That there can be no sound and complete deductive systematization of elementary number theory, much less of pure mathematics generally, is true. It is decidedly paradoxical, in the sense that it upsets crucial preconceptions. We used to think that mathematical truth consisted in provability.

Like any veridical paradox, this is one we can get used to, thereby gradually sapping its quality of paradox. But this one takes some sapping. And mathematical logicians are at it, most assiduously. Gödel's result started a trend of research that has grown in 30 years to the proportions of a big and busy branch of mathematics sometimes called proof theory, having to do with recursive functions and related matters, and embracing indeed a general abstract theory of machine computation. Of all the ways of paradoxes, perhaps the quaintest is their capacity on occasion to turn out to be so very much less frivolous than they look.

THE THEORY OF TYPES[1]

by Paul Weiss

It would seem from the interpretation that Whitehead and Russell put on the theory of types, that it is impossible or meaningless to state propositions which have an unrestricted possible range of values, or which, in any sense, are arguments to themselves. Thus on the acceptance of the principle that statements about all propositions are meaningless,[2] it would be illegitimate to say, "all propositions are representable by symbols," "all propositions involve judgment," "all propositions are elementary or not elementary," and if no statement could be made about all the members of a set,[3] it would be impossible to say, "all meanings are limited by a context," "all ideas are psychologically conditioned," "all significant assertions have grammatical structures," etc., all of which are intended to apply to themselves as well. The theory seems also to make ineffective a familiar form of refutation. General propositions are frequently denied because their enunciation or acknowledgment depends on the tacit supposition of the truth of a contradictory or contrary proposition. Such refutations assume that the general proposition should be capable of being an argument of the same type and to the same function as its own arguments, so that according to Whitehead and Russell, they fallaciously refute "by an argument which involves a vicious circle fallacy".[4]

That these limitations on the scope of assertions or on the validity of refutations are rarely heeded is apparent even from a cursory examination of philosophical writings since 1910. Thus Russell, apropos to Bergson's attempt to state a formula for the comic says,[5] "it would seem to be impossible to find any such formula as M. Bergson seeks. Every formula treats what is living as if it were mechanical, and is therefore by his own rules a fitting object of laughter." The characterisation of all formulæ, even though it refers to a totality, seems to Mr. Russell to be of the same type as the formulæ characterised.

[1] Chap. II., *Principia Mathematica.*
[2] P. 37, *ibid.* (second edition). [3] P. 37, *ibid.* [4] P. 38, *ibid.*
[5] "Prof. Guide to Laughter," *Cambridge Review*, Vol. 32, 1912, and Jourdain's *Philosophy of Mr. B*tr*nd R*ss*ll*, pp. 86-7.

If the theory were without any embarrassments of its own, and were indispensable for the resolution of the so-called paradoxes[1] (which no one seems to believe), there would be nothing to do but to acknowledge the impossibility of cosmic formulations, as well as the inadequacy of philosophic criticisms, and to pass charitably over such remarks as Russell's as mere accidents in a busy life. However, the statement of the theory itself involves the following difficulties in connection with (1) its scope, (2) its applicability to propositions made about it, and (3) its description.

1. It is either about all propositions or it is not.

A. If it were about all propositions it would violate the theory of types and be meaningless or self-contradictory.

B. If it were not about all propositions, it would not be universally applicable. To state it, its limitations of application would have to be specified. One cannot say that there is a different theory of types for each order of the hierarchy, for the proposition about the hierarchies introduces the difficulty over again.

2. Propositions about the theory of types (such as the present ones, as well as those in the *Principia*) are subject to the theory of types, or they are not.

A. If they were, the theory would include within its own scope propositions of a higher order, and thus be an argument to what is an argument to it.[2]

B. If they were not, there would be an unlimited number of propositions, not subject to the theory, that could be made directly or indirectly about it. Among these propositions there might be some which refer to a totality and involve functions which have arguments presupposing the function.

3. The statement of the theory of types is either a proposition or a propositional function, neither or both.

A. If it were a proposition, it would be either elementary, first order, general, etc., have a definite place in a hierarchy and refer only to those propositions which are of a lower order. If it were held to be a proposition of the last order, then the number of orders would have a last term, and there could not be meaningful propositions made about the theory. The *Principia* should not be able to say, on that basis, just what the purpose, character and application of the theory is.

B. Similarly, if it were a propositional function, it would have a definite place in a hierarchy, being derived from a proposition by generalisation. It could not refer to all propositions or propositional functions, but only to those of a lower order.

[1] Paradoxes, though contrary to common opinion, may be and frequently are true. Paranoumena, violating principles of logic or reason, if they are not meaningless, are false, and it is only they which are capable of logical analysis and resolution. What the *Principia* attempts to do is to solve apparent paranoumena by a paradox.

[2] P. 39, *Principia Mathematica*.

C. If it were neither it could not be true or false, nor refer to anything that was true or false. It could not apply to propositions, for only propositions or propositional functions, in a logic, refer to propositions.

D. If both at once, it would be necessarily self-reflexive.

a. If as function it had itself as value, it would refer to itself. But the theory of types denies that a function can have itself as value.

b. If as function it had something else as value, it would conform to the theory, which insists that functions have something else as values. The theory then applies to itself and is self-reflexive, and thus does not apply to itself. As, by hypothesis, it is a value of some other function, there must be propositions of a higher order and wider range than the theory of types.

It is no wonder that the perpetrators of the theory have not been altogether happy about it! What is sound in it—and there is much that is—is best discovered by forgetting their statements altogether, and by endeavouring to analyse the problems it was designed to answer, without recourse to their machinery. The result will be an acknowledgment of a theory of types having a limited application, and a formulation of a principle which will permit certain kinds of unrestricted general propositions.

To do this we shall deal in detail with two apparent paranoumena dealt with in the *Principia*, where the difficulty is largely *methodological*. We shall then treat of Weyl's "heterological-autological" problem, where the difficulty is due to a confusion in *meanings*. Those problems which cannot be dealt with under either heading will be those which need a theory of types for their resolution.

1. *Epimenides.* The proposition "All Cretans are liars" must be false if it applies to Epimenides as well, for it cannot be true, and only as false has it meaning. If it were true, it would involve its own falsity. When taken as false, no contradiction, or even paradox, is involved, for the truth would then be "*some* Cretans tell the truth". The truth could not be "all Cretans tell the truth" for Epimenides must be a liar for that to be true and by that token it must be false. Epimenides himself would be one of the lying Cretans, and one of the lies that the Cretans were wont to make would be "all Cretans are liars". Thus if Epimenides meant to include all his own remarks within the scope of the assertion, he would contradict himself or state a falsehood. If it be denied that a contradictory assertion can have meaning, he must be saying something false if he is saying anything significant. Had he meant to refer to all other Cretans there is, of course, no difficulty, for he then invokes a kind of theory of types by which he makes a remark not intended to apply to himself. All difficulty disappears when it is recognised that the formal implication, "all Cretanic statements are lies" can as a particular statement be taken as one of the values of the terms of this implication. Letting $Ep\,!\,p$ represent "Epimenides once

asserted p"; ϕ represent "Cretanic" and p represent a statement or proposition, then for "All Cretanic statements are false (or lies)," we have:

$$1.\ \phi p . \supset_p . \backsim p.$$

And as Epimenides is a Cretan, for any assertion he makes we have:

$$2.\ Ep\,!\,p . \supset_p . \phi p.$$

As $\#1$ is an argument to the above—it being Epimenides' present remark—we get:

$$3.\ Ep\,!\,\{\phi p . \supset_p . \backsim p\} . \supset . \phi\{\phi p . \supset_p . \backsim p\}$$

$\#1$, as a Cretanic statement, is an argument to $\#1$ as a formal implication or principle about Cretanic statements, so that:

$$3A.\ \phi\{\phi p . \supset_p \backsim p\} . \supset . \backsim\{\phi p . \supset_p . \backsim p\}$$

$\#3$ and $\#3A$ by the syllogism yield:

$$3B.\ Ep\,!\,\{\phi p . \supset_p . \backsim p\} . \supset . \backsim\{\phi p . \supset_p . \backsim p\}$$

so that in this instance Epimenides lied.

It is important to note that $\#1$ states a formal implication, and that $\#3$, $\#3A$ and $\#3B$ employ $\#1$ as a particular assertion or specific argument to their functions. $\#3A$ is an instance of the implicationexpressed by $\#1$, and is this instance because of the particular argument it does have. It states the fact that "'all Cretanic statements are false' is a Cretanic Statement," implies that "'all Cretanic statements are false' is false". Substitution of another argument would give a different instance; though of course of the same implication. The implication contained in its argument does not have instances. "'Some Cretanic statements are false' is a Cretanic statement" or "'This Cretanic statement is false' is a Cretanic statement" are not instances of "'All Cretanic statements are false' is a Cretanic statement," but of "P is a Cretanic statement". These three propositions have different subjects; they are different values of the same propositional function. That these subjects have relations to one another is of no moment. "My wife loves me" and "my mother-in-law is old (or loves me)" are two distinct and logically independent propositions, even though there is a relationship between the two subjects.

It is because any considered general proposition is at once an individual fact, and a formal implication or principle, with many possible arguments, that it is capable of being taken as an argument to itself. All propositions about words, logic, truth, meaning, ideas, etc., take arguments which fall in these same categories, and in so far as such a general proposition is stated in words, determined by logic, etc., it should, as such a fact, be an argument to itself as a formal implication. The principle must be false if this cannot be done, for it is sufficient, in order to overthrow a proposition of this kind, to produce one argument for which it does not hold. One may limit the principle by asserting that it holds for "all but . . .",

in which case it is a *restricted* general proposition. Nominalism, association of ideas, scepticism, the theory of universal tautology, the denial of logic are defended in propositions which cannot take themselves as arguments, and which as facts are arguments to contradictory principles. Their contradictory principles therefore hold sometimes at least, so that these doctrines must be false if they are put forward without restriction, and cannot be universally true, if, in Bradley's words, they "appear".

2. "I am lying"—if it be taken in isolation from all fact—is a meaningless statement. There must be some objective truth that is distorted, and unless it is provided the assertion has no significance. This proposition means either, "I am lying about X"; "I always lie," or "I have always lied". The first can be either true or false without giving rise to any problem, except where "all my assertions" is made an argument to X, in which case it is equivalent to either the second or third formulation. "I always lie" involves the same situation as with Epimenides, and the proposition is false. The supposition of its truth would involve a contradiction; the supposition of its falsity means simply that I sometimes lie and sometimes tell the truth. If what is meant is that "I have always lied" that does not involve a contradiction, for what is intended is a restricted proposition, applying to *all but* the present one. It can be true because it does not apply to all propositions; if it were false, then sometimes I lied and sometimes I did not. In short, there is nothing like a self-reflective universal liar, which is an interesting moral conclusion to derive from a logical analysis. Similarly, there cannot be a thorough scepticism held by the sceptic to be valid.

Prof. Whitehead (to whom I am also indebted for the notation) has pointed out to me that wherever a conjunction of propositions results ir a *reductio ad absurdum*, there is no way of determining on logical grounds alone which of the antecedents fails, or is false (though one at least must be). Thus in the case of Epimenides we have:

$$4.\ \underset{(A)}{\{\phi p \,.\supset_p.\, \backsim p\}} \,.\, \underset{(B)}{\{Ep\,!\,p \,.\supset_p.\, \phi p\}} \,.\, Ep\,!\,\underset{(C)}{\{\phi p \,.\supset_p.\, \backsim p\}} \,.\supset.\backsim \underset{(D)}{\{\phi p \,.\supset_p.\, \backsim p\}}$$

It is because B and C are in that case assumed to hold, that we can say that A must fail. If the truth of all these antecedents were undetermined, we should have merely the general rule: a *reductio ad absurdum* has as a necessary condition the conjunction of one or more false propositions. Transposition—

$$4'.\ \underset{(D)}{\{\phi p \,.\supset_p.\, \backsim p\}} \,.\supset.\backsim \underset{(A)}{\{\phi p \,.\supset_p.\, \backsim p\}} \,.\vee.\backsim \underset{(B)}{\{Ep\,!\,p \,.\supset_p.\, \phi p\}} \,.\vee.\backsim Ep\,!\,\underset{(C)}{\{\phi p \,.\supset_p.\, \backsim p\}}$$

makes it apparent that to deny the conclusion of a *reductio ad absurdum* is to imply that at least one of the antecedents is false.

In connection with the *reductio ad absurdum* involved in the assertions, "I always lie" and "I always doubt," #4B reduces to the tautologies: "If I assert p, p is my assertion," and "If I doubt, the doubt is mine". In these cases, the only alternatives left are the denial of the fact of the assertion (#4C), or the truth of the principle itself (#4A).

3. Weyl's heterological-autological contradiction[1] is the result of a material fallacy of amphiboly in connection with the employment of adjectives. The simplest form of such a fallacy is due to a failure to distinguish between an adjective as substantive and an adjective as attribute. Thus if we treat both the subject and attribute in "large is small" and "small is large" as attributes united by a copula expressing identity (instead of reading it as "large is a small word," "small is a large word") we could say "whatever is small is large, and whatever is large is small". No one, I believe, since the Megarics, has been troubled by this particular confusion.

The present problem is the result of a confusion, not between substantive and adjective, but between an adjective which expresses a property, and an adjective which expresses a relation between this property and the substantive. All words can be described in terms of a property—they are long, short, beautiful, melodious, etc., words. They can be classified in accordance with these properties, giving us the class of long words, short words, etc. They can also be classified as either "autological" or "heterological," depending on whether or not the same word is at once substantive and property-adjective; the terms "autological" and "heterological" expressing relationships between the substantive and adjective.

The autological class is made up of words, each of which expresses a property which it possesses; though all of them have unique properties. If "short" be short, and if "melodious" be melodious, they would both be members of the autological class; though in addition, "short" would be a member of the class of short words, and "melodious" would be a member of the class of melodious words.

The heterological class is made up of words, each of which expresses a property which it does not possess. If "long" be short, and if "fat" be thin, they would both be members of the heterological class; although here also "long" would be a member of the class of short words, and "fat" would be a member of the class of thin words. Though when classified according to the relationship of the adjective to the substantive, "short" would be an autological word and "long" a heterological word, they would both be members of that class which was defined in terms of the properties of words—being in this case, members of the class of short words.

[1] Briefly stated it is: all words which express a property they possess are autological; all words which express a property they do not possess are heterological. If 'heterological' is heterological it expresses a property it possesses and is thus autological; if it is autological, it expresses a property it does not possess and is therefore heterological. *Das Kontinuum*, p. 2.

Now if heterologicality were a property that a word could have, and if the word "heterological" had that property, it would be a member of the autological class, for it would then possess a property that it expressed. But it would also be a member of a class of words which had the *property* of heterologicality. This class is determined by taking the properties of words, and if it be called "heterological," must be distinguished from that class which was determined not by properties, but by the relationship between properties and substantives.

If there were a property like autologicality and if "heterological" had that property,[1] it would be a member of the heterological class, for it would express a property which it did not possess. But it would also be a member of the class of words which possessed autologicality and could be thus classified.

Thus if "heterological" had the property of autologicality, it would be in the heterological class owing to the *relation* which held between the property and substantive (or between a property it possessed and the property it expressed); but it would be in the class of autological words, owing to a *property* it possessed. If it had the property of heterologicality, it would be in the autological class on the basis of the *relation*, and in the class of heterological words on the basis of *property* classification. There is no difficulty in considering something as a member of two distinct classes, owing to the employment of different methods of classification. There is no contradiction in saying: "'heterological' expresses the property heterologicality, possesses the property autologicality, and the relation between these properties is heterological, in that it expresses and possesses the property heterologicality and the relation between them is autological." Similarly, Richard's contradiction, Berry's contradiction, and that involving the least indefinable ordinal, are resolvable by recognising that "nameable" and "indefinable" are used in two sharply distinguishable senses. They do not require a hierarchy, but a discrimination in the methods of description.

When a distinction is made between a class and its membership (the distinction between a number of numbers and a number is a particular case of this), and between a relation of objects and a relation of relations, the requirements for the solution of the other mathematical problems are provided. A class is other than its members, and a relation, like all universals, transcends any given instance or totality of instances. As they have characters of their own, universals can be described in terms of other universals, which in turn transcend them. Arguments are of a different "type" than functions, just so far as they have different logical characteristics,

[1] 'Heterological,' in fact, has the properties of being long, polysyllabic, etc., and it is questionable whether there are properties like autologicality and heterologicality possessed by words. If there be no such properties, "heterological" is a member of the class of long words, polysyllabic words, etc. In addition it would be one of the terms related by the heterological relation, which fact would not make it have the *property* of heterologicality.

i.e. are different kinds of logical facts. The class which is an argument to a function about classes has, as argument, a different logical import than the function, and its arguments have a different import from it. This is true of all functions, restricted and unrestricted alike, for it means simply that they are discriminable from their arguments. They can, despite this difference, have characteristics in common with their arguments, and are to that extent unrestricted. Thus in the case of "the class of those classes which are identical with themselves," the class of classes can be taken simply as a class, without logical embarrassment. Yet a class of classes differs from a class, and must therefore be capable of a different characterisation, and thus also be an argument to a function of a different type. With some classes, it may not be possible to consider them as arguments to their own functions, without uncovering a contradiction. In such cases (*e.g.* the class of those classes which are not members of themselves, and the relations which are connected by their contradictories), it is the difference between the function and the argument that is of moment. That *some* cannot take themselves as arguments does not indicate that all classes or functions are restricted in scope, but simply that classes and functions are *non-restricted*. Some classes and functions are restricted and some are not. To say that all are because some are is an obvious fallacy.

Whenever, as individual, a general proposition is in the class of those objects of which it treats, but cannot be considered as an argument to itself, it is either false or restricted in scope. If the second, its range of arguments must be specified. Accordingly, we can state as a *necessary* condition for the truth of a general proposition, whose scope is unspecified, that when it has a character, which is one of the characters about which it speaks, it *must* be an argument to itself. Thus if Bergson adequately described the comic, his formula should be an object of laughter, and if the theory of types is universal in application, it should be capable of being subject to itself. Conformity to this condition indicates that the unrestricted proposition is *possibly* true; not that it is necessarily true. To demonstrate that such a proposition was necessarily true, it would be essential to show that the supposition of its falsity assumes its truth. That there is danger in applying this rule can be seen from the consideration of some such proposition as: "Everything is made up of language elements". Its denial will be made up of language elements, and would seem to demonstrate that the proposition was necessarily true. Supposition of the falsity of a proposition, however, means verbal denial only in so far as the proposition applies to the realm of language. If it applies to everything, supposition of its falsity involves the positing of the objects of assertions; not the assertions. A necessary unrestricted proposition about everything can be supported only by a demonstration that the supposition of an argument for which it does not hold is self-contradictory. If the proposition has to do with grammar, meaning, logic, judgment, etc.,

the conditions for a necessarily true and unrestricted proposition would be: 1. the assertion of it is an argument to it; 2. any possible denial is an argument to it. That "any possible denial" rather than "any given denial" is required, is apparent from the consideration of the following propositions: "All sentences are made up of eight words," "No sentence is made up of eight words". Each of these contains eight words. It is because of the fact that we can formulate propositions such as, "It is false that every proposition must be made up of eight words," that the condition is seen not to have been met.

An unrestricted proposition applies to every member of the category, and has some aspect of itself as value. It is in some sense then a determinate in the category which it determines. If the proposition refers to some other category than the one to which it as fact, or some aspect of it as fact, belongs, it is restricted. Thus "all men are mortal" is neither man nor mortal, and as condition does not determine itself as fact. Any proposition referring to that statement would be of a different type, and would deal with its truth, falsity, constituents, historical place, logical structure, etc. Though the unrestricted propositions have no limitations, the category to which they refer may have. Epimenides' remark, for example, referred only to Cretans. As his assertion was a determinate in the category, and as his statement of the supposed conditions imposed on the members of that category was not a possible argument to the general proposition, the general proposition was seen to be false or restricted. Had he said, "All Cretans tell the truth," he would have stated an unrestricted proposition which was possibly true. It could not be said to be necessarily true unless Cretan and lie, against the evidence of history, were actually contradictories.

Accordingly, we shall say: *All true unrestricted propositions are arguments to themselves; or by transposition, those propositions which are not arguments to themselves are either restricted or false.* As this proposition can take itself as argument it is possibly true. Unless no proposition is possible which does not conform to it, it cannot be said to be necessarily true. I have not been able to demonstrate this and therefore accept it as a definition or "methodological principle of validation". The theory of types, in its most general form, may be stated as: *A proposition or function of order n, which cannot be an argument to itself, is, as fact, an argument of a proposition or function of order n + 1.*

In accordance with the scheme of the criticism of the theory of types, we can describe our principle as (1) applying to all propositions, including (2) those which refer to it. (3) It is a formal implication with itself as one of its arguments. The theory of types, on the other hand, (1) does not apply to all propositions, but only to those which are restricted, (2) *may* apply to those propositions which refer to it, and (3) is a formal implication which cannot take itself as argument.

The theory of types cannot be an unrestricted proposition about

all restricted propositions. As an unrestricted proposition it must take itself as argument; but its arguments are only those propositions which are *not* arguments to themselves. It cannot therefore be unrestricted without being restricted. Nor can it be a restricted proposition about all restricted propositions for it would then be one of the restricted propositions, and would have to take itself as argument—in which case it would be unrestricted. Hence it cannot be restricted without being unrestricted. Three possible solutions may be advanced. The first is that the theory of types is restricted and does not apply to *all* restricted propositions, but only to *some* of them. It is not an argument to itself but to some other proposition about restricted propositions. This in turn will have to be restricted and refer only to some propositions, and so on, giving us theories of types of various orders. The proposition made about the totality of these orders would be of a still higher order and would in turn presuppose a higher order *ad infinitum*. The theory of types thus depends on theories of types of theories of types without end. This seems probable on the ground that the theory is based on the recognition that no proposition can be made about all restricted propositions, so that it must by that very fact admit that it cannot apply to all of them. Instead, therefore, of the theory of types applying to all propositions, and determining them in various orders, it does not even apply to all of a given class of them. This interpretation would not affect unrestricted propositions, and would merely show that the determination of restricted propositions is subject to determinations without end.

The second possibility is suggested by the consideration of a proposition such as: "all truths are but partially true". If that were absolutely true, it would contradict itself, and if it were not, could only apply to some truths. Considered as referring to the necessary limitations which any finite statement must have, it would take itself as argument in so far as it was finite, thus indicating that it was absolutely true about finite propositions, and yet not absolutely true as regards all truths. By pointing out the limitations of a finite statement it indicates that there is an absolute truth in terms of which it is relatively true. On this interpretation, any condition which imposes universal limitations is unlimited in terms of what it limits, but limited in turn by some other condition. One might hold, therefore, that the theory would be unrestricted as regards restricted propositions, and restricted as regards all propositions, and would point to a higher principle which limits it.

The third possibility is to allow for "intensive" propositions which are neither restricted nor unrestricted, being incapable of any arguments. The theory of types could be viewed as such an intensive proposition, and what we have called its arguments, would merely "conform" to it. This interpretation means the downfall of a completely extensional logic, and a determination of an extensional logic as subordinate to an intensional one.

There are difficulties in each of these interpretations. We shall

not now choose among them. In any of these cases, however, a restricted proposition which refers to some other than the restricted aspect of the theory would be subject to the theory and the principle we have laid down about unrestricted propositions could still hold. Those restricted propositions which refer to the restricted character of the theory would not be an argument to it on the first, would be an argument to it on the second, and would neither be nor not be an argument to it on the third solution.

To briefly summarise: The theory of types must be limited in application. Not all the problems it was designed to answer require it; another principle of greater logical import is desirable; while for the resolution of the problems in which it is itself involved, very drastic remedies are necessary. No matter how the theory fares, the possibility of the methodological principle and the possibility of other solutions for the so-called paradoxes, indicate that it is at least not as significant an instrument as it was originally thought to be.

PAUL WEISS.

A system which can define its own truth.

By

John Myhill (Philadelphia).

Tarski has shown [1]) that for a certain class of logical systems S. the following holds:

It is impossible to define in S the class of Gödel-numbers of true statements of S.

The essence of his proof consists of the following version of the Epimenides. Let *Fmla* be the class of Gödel-numbers of meaningful statements of S; then we can define the class of false statements of S as follows

$$x \in Fals \equiv x \in Fmla \cdot \sim x \in True$$

where *True* is the class of Gödel-numbers of true statements of S.

Let „$Nom\,(y,x)$" say that y is the Gödel-number of the numeral designating x, and let „$Subst\,(z,y,x)$" say that z is the Gödel-number of the result of writing the expression whose Gödel-number is y for all free occurrences of „v" in the expression whose Gödel-number si x. Let n be the numeral designating the Gödel-number of

Ep. 1 $(Ey)\,(Ez)\,(Nom\,(y,v) \cdot Subst\,(z,y,v) \cdot z \in Fals)$.

Then the formula

Ep. 2 $(Ey)\,(Ez)\,(Nom\,(y,n) \cdot Subst\,(z,y,n) \cdot z \in Fals)$

says that the result of writing n for all free occurrences of „v" in Ep. 1 is false. But this result is Ep. 2 itself; i. e. Ep. 2 affirms its own falsehood, an evident contradiction [2]).

---

[1]) A. Tarski, *Pojęcie prawdy w językach nauk dedukcyjnych*, Warszawa, 1933.

[2]) We have $(Ey)\,(Ez)\,(Nom\,(y,n) \cdot Subst\,(z,y,n) \cdot z \in Fals) \equiv \neq (Ep.\ 2\ is\ true)$; but by Tarski's schema for truth (see Tarski, op. cit.), also $(Ey)\,(Ez)\,(Nom\,(y,n) \cdot Subst\,(z,y,n) \cdot z \in Fals) \equiv (Ep.\ 2\ is\ true)$; the contradiction follows by the theory of deduction.

It is obvious that this proof of Tarski's depends upon S's containing a certain amount of conceptual apparatus; in particular it depends upon S's containing negation. The question has hitherto remained undecided, whether *any* system, even without negation, can define its own truth. The purpose of this paper is to answer this question in the affirmative.

Rósza Péter[3]) has constructed a number-theoretic function Φ such that for every primitive recursive function f of two arguments there is a number x such that

$$f(y,z)=\Phi(x,y,z)$$

for all y and z. Further, it is evident from the definition of this function that it is general recursive.

Let S_1 be a system consisting of the recursion equations for Φ and everything which can be deduced from them by the use of extensionality and substitution of constants for variables.

Let S_2 be the class of all formulae of S_1 which contain no free variables.

Let S_3 be a system consisting of all formulae of S_2 and everything which can be obtained from them by means of the rule:

From „$\ldots n\_$“, where „n“ is a numeral, infer „$(Ex)(\ldots x\_)$“, where „x“ is a variable not occurring in „$\ldots n\_$“.

We shall show that S_3 *can define its own truth.*

It is evident that S_3 is a system, i. e. that the class of Gödel-numbers of theorems of S_3 forms the range of values of a general recursive function, say α. Further S_2, and hence S_3, is clearly complete and consistent, in the sense that all true formulae expressible in the notation of S_2 and S_3, and no others, are provable in S_2 and S_3 respectively. Hence the class of true statements of S_3 coincides with the class of theorems of S_3.

Rosser[4]) has shown that the range of values of every general recursive function coincides with the range of values of some primitive

---

[3]) R. Péter, *Konstruktion nichtrekursiver Funktionen*, Math. Ann., **111** (1935), pp. 42-60.

[4]) B. Rosser, *Extensions of Some Theorems of Gödel and Church*, Journal of Symbolic Logic, **1**, p. 88, Lemma I, Corollary 1.

recursive function. Hence we may suppose α primitive recursive. We have

$$\begin{aligned} x \text{ is the Gödel-number of a true statement of } \mathcal{S}_3 &\equiv (Ey)\,(x = \alpha(y)) \\ &\equiv (Ey)\,(x = \beta(y,0)) \\ &\equiv (Ey)\,(x = \Phi(m,y,0)) \end{aligned}$$

for some primitive recursive β and for some m and this is clearly expressible in $\mathcal{S}_3$; hence $\mathcal{S}_3$ can define its own truth.

Q. E. D.

HETEROLOGICALITY

By GILBERT RYLE

IN ANALYSIS (N.S. No. 16, March 1950) Mr. Lawrence ably criticises Russell's use of an alleged paradox (that of' heterological ' being homological, if heterological, and *vice versa*) as a proof of the thesis of language-hierarchies. I agree with his conclusions, but I do not think he brings out the main reason why it is an improper question even to ask whether ' heterological ' is heterological or homological.

Some people introduce themselves on meeting strangers.

[Continued on next page]

Others do not do this. They might be distinguished as ' self-nominators ' and ' non-self-nominators '. If Dr. John Jones, on meeting a stranger begins by saying ' Dr. John Jones ' he acts as a self-nominator. So, too, if he says ' Captain Tom Smith', save that now he gives a pseudonym. If he says ' it's a fine day ', ' I'm a doctor ' or ' I'm a self-nominator (or a non-self-nominator) ', then he has given a true or false description of the weather, his profession or his practice of announcing or withholding his name ; and in the last case his action belies or bears out his remark. But he has not given his surname, Christian name, nickname or pseudonym. ' . . . a self-nominator ' and ' . . . a non-self-nominator ' are the tail ends of character descriptions. Such descriptions require that the person described has a name to give, though this is only referred to, not given. Epithets are not names, even when they carry references to names.

A perverse parent might name his child ' Non-self-nominator ', as there was once a club named ' The Innominate Club '. The son might then correctly give his name as ' Master Non-self-nominator Brown '. But he would no more have belied his announcement than a big man does by introducing himself as 'Alderman Little '. Names are not epithets, even though the same vocables constitute an epithet in some language or other.

To say ' I am a self-nominator (or non-self-nominator) ' not only is not to give one's name ; it is not even to describe one's name, as I might describe my name by saying that it rhymes with ' smile '. Unfortunately the term of art ' description ' is often used to cover idioms which, in ordinary life, would not be called ' descriptions ' at all. No policeman would accept as a description the phrase ' my name ' (in ' I never volunteer my name '). What I refer to with such expressions can, in principle, be specified or particularised. But ways of referring to things are not ways of answering consequential requests for information about those things. Very often we employ referring expressions, which also have a descriptive force, e.g. ' she ' and ' yonder lame horse ', and a hearer may get the reference while contesting the attribution of femininity or lameness.

For it to be true that I do or do not give my name, there must be a ' namely-rider ' of the pattern " . . . , namely ' John Jones ' ". Nor could it be true that a person had a profession or a disease without there being a namely-rider, of the pattern " . . . namely, the Law " or " . . . namely, asthma ". The namely-riders need not be known to the persons who make or understand the statements that have them. I know that there is a day of the week on

which I shall die, but I do not know what it is. The day of my death could not be *just* the day of my death, but no particular day of the week.

Like anything else, linguistic expressions can be described or misdescribed by appropriate epithets. An epitaph, a peroration, a sonnet, a phrase, or a word may be English, mispronounced, vernacular, full of sibilants or obscene. The vocabulary of philologists is necessarily full of epithets appropriate to linguistic expressions, as the vocabulary of botanists is of those appropriate to plants. The methods of finding out what epithets can be truly applied to particular expressions are (crude or refined) philological methods. The properties of expressions established by these methods are philological properties. Thus 'dissyllabic', 'of Latin origin', "rhymes with 'fever'", 'solecism' and 'guttural' are philological epithets, standing for philological properties.

Most epithets, like astronomical, pharmaceutical and botanical epithets, are not appropriate to linguistic expressions. Cottages, but not phrases, may be thatched or unthatched, tiled or untiled. Reptiles do or do not hibernate; adverbs neither do nor do not. When an epithet is not a philological epithet, the question whether or not it applies to a given expression is an improper question.

Philological epithets do not constitute a proper genus. There are widely different interests that amateur or professional philologists may take in expressions, and widely different questions that they can raise about them. Phonetic epithets are not congeners of stylistic, orthographic, grammatical or etymological epithets.

It is no accident that a philological epithet may stand for a property of expressions and itself be one of the expressions that has that property. 'English', like 'cow', is an English word, 'polysyllable', like 'Saturday', is a polysyllable. Conversely, a philological epithet may stand for a property the possession of which can be truly denied of that epithet. 'French' is not a French word; 'monosyllable' is not a monosyllable; 'obsolete' is not obsolete. The question whether the property for which a philological epithet stands does or does not belong to that epithet itself is a proper question, e.g. whether 'dactyl' is or is not a dactyl and whether 'mispelt' is or is not misspelt.

One may then construct two artificial parcels, one of those philological epithets which have the philological properties for which they stand; the other of those which lack the properties for which they stand. The first might be called 'self-epithets';

the second ' non-self-epithets '. ' Deutsch ' is a self-epithet, ' obscene ' a non-self-epithet. ' Written,' here, is a non-self-epithet ; in manuscript it was a self-epithet.

Now the question is supposed to arise "Are ' self-epithet ' and ' non-self-epithet ' themselves self-epithets or non-self-epithets? In particular, is ' non-self-epithet ' a non-self-epithet or a self-epithet? "

But before starting to wriggle between the horns of this dilemma, we must consider whether the question itself is a proper question ; i.e. whether ' self-epithet ' and ' non-self-epithet ' belong to the class of philological epithets at all. As we have seen, most epithets do not belong to this class, so we cannot assume that these two (invented) classificatory expressions do so either. Certainly at first sight they look as if they do belong to it, since we can ask whether ' monosyllabic ' is a self-epithet or not, and decide that it is a non-self-epithet. The predicate end of the sentence " ' monosyllabic ' is a non-self-epithet " does look like a philological epithet of ' monosyllabic '. It's a predicate of a quoted expression, isn't it ?

But, if we recollect that ' self-nominator ' and ' non-self-nominator ' were not names but epithets carrying references to unmentioned names, we may suspect that ' self-epithet ' and ' non-self-epithet ' are not philological epithets, but expressions carrying references to unmentioned philological epithets. And this suspicion can be confirmed.

How do we decide (*a*) whether words (including ' English ' and ' polysyllabic ') are English or polysyllabic, etc., and (*b*) whether words, like ' English ' and ' polysyllabic ', are or are not self-epithets? We decide questions of the former sort by (elementary or advanced) philological methods. We consult dictionaries, count syllables, use stop watches, and listen to linguaphone records. But we do not decide whether words are self-epithets or not by examining them by further philological methods. There is no visible, audible, grammatical, etymological or orthographical feature shared by ' French ', ' monosyllabic ', ' obscene ' and ' misspelt '. On the contrary, we call them ' non-self-epithets ' because we find by various philological methods that ' French ' is not a French word, ' obscene ' is not an obscene word, ' misspelt ' is not misspelt ; and we have decided to list as ' non-self-epithets ' all those philological epithets which lack the philological properties for which they stand. We find by philological methods that these and other quite different philological epithets stand severally for, perhaps, not even generically kindred philological properties, and that

the epithets lack the properties for which they stand; we then collect them together as ' non-self-epithets ' not in virtue of any common philological properties that they all stand for but in virtue of their being alike in lacking the philological properties for which they severally stand. So ' non-self-epithet ' and ' self-epithet ' do not stand for any philological properties of expressions. Their use is to assert or deny the possession of ordinary philological properties by the epithets which stand for them. Their use *presupposes* the ordinary use of philological epithets in the description of expressions of all sorts (including philological epithets); it is not, therefore, a part of that ordinary use. ' Orthographic ' is a self-epithet only *because* ' orthographic ' both has and stands for a certain philological property, namely that of being correctly spelled. Unless there was an opening for this namely-rider, there would be no job for the expression ' self-epithet ', just as there would be no job for the expression ' self-nominator ' if there were no opening for such a namely-rider as " . . . namely, ' John Jones ' ". As ' self-nominator ' carries a reference to a name which it is not and does not mention, so ' self-epithet ' carries a reference to an ordinary philological property which it does not itself stand for.

At first sight, however, the statement that ' English ' is a self-epithet does seem not merely to refer to but to mention the philological property of being in the English tongue. For we English speakers are all so familiar with the use of the word ' English ', that we are tempted to suppose that to be told that ' English ' is a self-epithet is to be told that ' English ' is an English word. But this is not so. A person who knew no German but had been told the use of ' self-epithet ' and ' non-self-epithet ' could understand and believe the statement that the German adjective ' lateinisch ' is a non-self-epithet. He would be believing that ' lateinisch ' stands for some specific philological property or other (he would not know which) and that ' lateinisch ' lacks that property. He knows that it is now proper to ask the rider-question ' Namely, which philological property? ' since if this question had no answer, then ' lateinisch ' could not be either a self-epithet or a non-self-epithet. For ' lateinisch ' to be a non-self-epithet, it must be false that ' lateinisch ' is *lateinisch*, whatever being *lateinisch* is. ' Lateinisch ' could no more be *just* a non-self-epithet, without there being a specific, but as yet unmentioned, philological property which it lacked, than a person could be *just* christened, without being christened ' John ' or ' James ' or . . .

' Self-epithet ' and ' non-self-epithet ' do not stand for any

of the specific philological properties which could be mentioned in the namely-riders for which they leave openings, any more than " christened ' . . . . ' " is a Christian name. Logicians' category-words are not among the words listed under those category-words. ' Fiction ' is not the name of one of the novels catalogued under the librarian's heading of ' Fiction '. ' Self-epithet ' and ' non-self-epithet ' are not philologists' epithets, but logicians' ways of dividing philologists' epithets into two artificial parcels, which, naturally, are not among the contents of those parcels.

In this connection I adopt and adapt an important argument of Mr. Lawrence (p. 78). ' English ' is a self-epithet, because ' English ' is an English word. But ' self-epithet ' does not *mean* . . . ' is an English word '— else " ' Deutsch ' is a self-epithet " would be false, instead of being true. Or if " ' English ' is a self-epithet " could be paraphrased by " ' English ' is an English word ", then since German words are not English and ' Deutsch ' is a German word, " ' Deutsch ' is a non-self-epithet ' would have to be true, when in fact it is false. Of itself, ' self-epithet ' tells us nothing about the tongues to which ' English ' and ' Deutsch ' belong, any more than about the tongue to which ' polysyllabic ' belongs, which is also a self-epithet. ' Either English or German or Lithuanian or a self-epithet " would be an absurd disjunction. So would any disjunction be absurd in which ordinary philological epithets were coupled by ' or ' with ' self-epithet ' or ' non-self-epithet '. ' Self-epithet ' and ' non-self-epithet ' convey no philological information about words. They are specially fabricated instruments for talking *en bloc* about the possession or non-possession by philological epithets of whatever may be the philological properties for which they stand. Such instruments are not among the philological epithets that they help logicians to discuss.

To put all this in the official terminology of ' heterological ' and ' homological '. We can say " ' obsolete ' is heterological," because ' obsolete ' has not gone out of fashion and ' obsolete ' *means* ' gone out of fashion '. We can say " ' polysyllabic ' is homological," because ' polysyllabic ' *is* a word of many syllables and *means* ' of many syllables '.

Now the words ' heterological ' and ' homological ' have and lack a number of ordinary philological properties. They are adjectives, polysyllabic, English (perhaps), cacophonous, aspirated and neologistic ; they are *not* prepositions, slang, or of Latin origin. But what we are asked to decide is whether

' heterological ' and ' homological ' are themselves heterological or homological, i.e. whether among the philological properties which ' heterological ' and ' homological ' have and lack, they have or lack the philological properties of homologicality and heterologicality. But to ask this is to suppose that ' heterological' and ' homological ' *do* stand for philological properties, i.e. that there could be words which were *just* heterological or *just* homological and not heterological or homological *because* lacking or possessing such and such ordinary philological properties. And this supposition is false. For to say that a word is heterological or homological is to refer to, without mentioning, some philological property (*not* heterologicality or homologicality) for which that word stands and which does or does not belong to that word. In using ' heterological ' and ' homological ' we are not mentioning word-properties, but referring to unmentioned word-properties. And references to unmentioned word-properties are not mentions of those or of extra word-properties, any more than references to unspecified diseases are themselves the specifications of those or of extra diseases. If unpacked, the assertion that ' heterological ' is heterological would run :—" ' Heterological ' lacks the property for which it stands, namely that of lacking the property for which it stands, namely that of lacking the property . . . " No property is ever mentioned, so the seeming reference to such a property is spurious.

This seeming paradox arises from treating certain umbrella-words, coined for the purpose of collecting epithets into two families as if they were themselves members of those families. To ask whether ' heterological ' and ' homological ' are heterological or homological is rather like asking whether Man is a tall man or a short man. We do not need stature-hierarchies to save us from deciding whether Man is a tall or a short man ; or language-hierarchies to save us from deciding whether ' heterological ' is heterological or homological. It is not even, save *per accidens*, a matter of minding our inverted commas but a matter of minding our grammar. Minding our inverted commas, in the required ways, *is* minding our grammar.

The same inattention to grammar is the source of such paradoxes as ' the Liar ', ' the Class of Classes . . . ' and ' Impredicability '. When we ordinarily say ' That statement is false ', what we say promises a namely-rider, e.g. ' . . . namely that to-day is Tuesday '. When we say ' The current statement is false ' we are pretending *either* that no namely-rider is to be asked for *or* that the namely-rider is ' . . . namely that the present

statement is false'. If no namely-rider is to be asked for, then 'The current statement' does not refer to any statement. It is like saying 'He is asthmatic' while disallowing the question 'Who?' If, alternatively, it is pretended that there is indeed the namely-rider, '. . . namely, that the current statement is false', the promise is met by an echo of that promise. If unpacked, our pretended assertion would run 'The current statement {namely, that the current statement [namely that the current statement (namely that the current statement . . .'. The brackets are never closed; no verb is ever reached; no statement of which we can even ask whether it is true or false is ever adduced.

Certainly there exist genuine hierarchies. My Omnibus 'Short Stories of O. Henry' is not a short story by O. Henry; nor is my batch of (five or six) Omnibus books a sixth or seventh Omnibus book. But 'I possess a batch of Omnibus books' has the namely-rider ". . . namely, 'The Short Stories of O. Henry', 'The Short Stories of W. W. Jacobs', etc., etc."; just as "I possess the Omnibus 'Short Stories of O. Henry'" has the namely-rider ". . . namely, 'The Third Ingredient', etc., etc." Obviously, we could go on to talk of the set of batches of Omnibus books in the possession of Oxford professors, and so on indefinitely. But at no stage can the mention of a set, or batch, or Omnibus *be* the mention of what it carries a reference to, but is not a mention of, namely what would be mentioned in their promised namely-riders, if these were actually provided.

In ordinary life, we commonly have different words with which we make our different references, such as 'Centre Forward', 'team', 'Division', 'League'. This prevents there being any linguistic temptation to talk as if a team might sprain its left ankle, or as if the League might win or lose the Cup. If we stick to one such word the whole way up the ladder (like 'class' and 'number') we are so tempted. Having failed to prevent such confusions arising, we have to seek remedies for them after they have arisen.

There are genuine hierarchies, so there exist (efficient and inefficient) ways of talking about them. But our (quite efficient) way of talking about footballers, teams, Divisions and Leagues is not itself to be described as a hierarchy of languages or ways of talking. Talking about the Berkshire League is not talking about talking about talking about the individuals who play, e.g. for a given village. Certainly the talk about the League is on a high rung of generality; a whole échelon of namely-riders goes

unstated. But no inverted commas are improvidently omitted. The relation of referring expressions to their namely-riders is not to be elucidated in terms of the relation of comments on expressions to the expressions on which they are comments. On the contrary, the relation of a comment to the expression on which it is a comment is (when it refers to this without citing it) just another case of the relation of a referring expression to the content of its namely-rider.

Many of the Paradoxes have to do with such things as statements about statements and epithets of epithets. So quotation-marks have to be employed. But the mishandling which generates the apparent antinomies consists not in mishandling quotation-marks but in treating referring expressions as fillings of their own namely-riders.

A team neither does nor does not sprain its left ankle. It is not one of its eleven possessors of left ankles, though a mention of the team carries a reference to these possessors of ankles. 'Heterological' neither has nor lacks any philological property for which it stands. It is not one of the philological epithets to which it and its opposite number are special ways of referring.

Magdalen College, Oxford.

VOL. LXII. No. 247.] July, 1953

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—SOME REFLECTIONS ON REFLEXIVITY

BY JØRGEN JØRGENSEN

IN this paper I wish to advance the view that so-called reflexive phenomena does not exist. Especially, I wish to defend the following theses :

(i) There are no reflexive relations or any other so-called reflexive phenomena.

(ii) In particular, no language or linguistic expression is self-referring.

(iii) The paradoxes of the Epimenides-type can be solved and explained in a quite elementary way without any need for a more or less complicated theory of logical types.

(iv) A simple form of such a theory seems necessary, however, in order to explain the syntactical use of the word " all ". By means of this theory various paradoxes, *e.g.* Russell's paradox, are eliminated and explained.

My conscious motives for trying to defend these theses are as follows :

First, I have always felt a certain disproportion existing between the complicated means used to avoid the paradoxes and the relative simplicity of the paradoxes themselves. It is generally thought foolish to dynamite butterflies, and as compared with the simpler paradoxes at least, the various theories of logical types seem to me to be of a too subtle character to be accepted lightly.

Secondly, the so-called solutions of the paradoxes have always left me in a certain state of uneasiness. True, the theories of types have, if tenable, shown how the paradoxes can be avoided, but they have not shown how they could arise. A quite satisfactory solution of them should, in my opinion, not only consist in directions as to how to avoid them, but should also expose the errors which cause them, *i.e.* it should not only prevent, but also explain them.

I think I can give an explanation which will serve as a means of preventing them, and which is at the same time much simpler than the theories commonly used for this purpose.

Let me start by a very elementary and well-known example of a logical paradox. The expression "This sentence is false" is generally considered paradoxical because it leads to contradictory conclusions: If it is true, then it is false; and if it is false, then it is true. Since these consequences are contradictory, the expression mentioned is considered meaningless. But it is likewise thought that the expression can obtain a meaning, if the so-called "systemic ambiguity" of the words "true" and "false" is taken into account. This I think is a mistake. The expression "This sentence is false" is not meaningless because it leads to contradictory consequences, but for a much deeper and simpler reason; and it cannot obtain a meaning by means of any logical theories or devices whatever.

My point is, that the expression "This sentence is false" is not a sentence at all in any logical sense. If the expression mentioned is considered a sentence and the word "false" is considered the logical predicate of this sentence, then the logical subject cannot be the whole expression, but at most either the words "This sentence" or the designation of these words. In the first case the whole expression is meaningless because "false" is not the kind of predicate which can be meaningfully ascribed to descriptions as "This sentence". And in the latter case the whole expression is meaningless because the description "This sentence" has no object, there being no sentence to which these words can refer. In neither case can any conclusions be drawn from the expression mentioned, and no paradoxes emerge.

If the whole expression "This sentence is false" is assumed truly to *be* a sentence affirming its own falseness, then it has no predicate, but the expression as a whole is functioning as the logical subject of a new sentence that must read: "The sentence 'This sentence is false' is false". Then the included sentence is true, and the including sentence false. And that is the end of it. No paradox emerges.

The included expression is meaningless, however, if it is considered to affirm its own falseness. This is seen when we remember that " false " is a predicate of sentences (or propositions). When it is predicated it, therefore, presupposes that there is a sentence of which it can be predicated (wrongly or truly). But if we are merely confronted with the expression " This sentence is false ", then there is no sentence of which " false " can be predicated.

The same applies, of course, to the expression " This sentence is true " which is also meaningless, although nobody has, to my knowledge, drawn any paradoxical consequences from it. And the same is, of course, the case with the expression " This sentence consists of six words ", which is neither true nor false, but simply meaningless, because there is no sentence to which it can refer. When the expression " This sentence consists of six words " is often considered a true sentence the reason is, in my opinion, that we are accustomed to conceive the description " This sentence " as a description of a sentence, and the expression mentioned being a sentence in a *grammatical*, although not in a *logical*, sense, we conceive this expression as the very sentence spoken about. But then our conception is really not expressed in the grammatical sentence mentioned, but must be formulated in the more complicated sentence: " The (grammatical) sentence ' This sentence consists of six words ' consists of six words ", where the included expression is the logical subject and the last occurring phrase " consists of six words " is the logical predicate.

It seems to me that this analysis also explains the paradox from which we started. In the expression " This sentence is false " the description " This sentence " is rightly assumed to refer to a sentence, and no other (grammatical) sentence than the expression mentioned being presented, it is wrongly assumed that the description refers to this very expression, which then loses its logical predicate and therefore is no logical sentence at all. Not observing this change because we seem to continue to see a sentence before us, we think that we can draw contradictory consequences from it and therefore believe that we get a paradox.

The most important generalization that can be made from these considerations is, to my mind, that no sentence can refer to itself, or that no sentences are self-referring. And this can, as far as I can see, be extended to any linguistic expressions whatever, nay, to all so-called reflexive phenomena. My reasons for believing this are as follows:

A phenomenon of whatever kind it may be *is* not a symbol, *i.e.* is not *functioning* as a symbol, unless it is contained in a process of symbolization as a part of this process taking place in an organism. The linguistic sounds or figures are in themselves or isolated from a process of symbolization not symbols, but solely sensational phenomena produced by symbolizing human beings for whom they may represent or refer to something different from themselves. Outside the symbolizing process they are but *remnants* of a linguistic process—just as a dead organism is but the remnants of a living organism. The sounds or figures are becoming symbols again, only if they are again used as such by becoming part of a new symbolizing process. A sentential expression as such, *i.e.* as an auditive or a visual phenomenon is not by itself referring to anything, but a *human being* can use it for referring to something different from the sentential expression. He cannot, however, by means of a sentential expression refer to this very sentential expression—this fact being a consequence of the very nature of symbolization. And in the same way a word can only function as a word, if it is part of a symbolizing process. If a word or a sentential expression is to function as a linguistic expression there must be something different from them to which they may refer *via* an organism, but they can never refer to themselves—that would be tantamount to their referring to nothing. Therefore, linguistic phenomena are never self-referring or reflexive.

And this applies, I think, to all other phenomena as well. There are no reflexive phenomena at all. This seems to me to be a simple consequence of the notion of relation. Any relation presupposes at least two terms which may be more or less alike in various respects, but which can never coalesce into a single term, if the relation shall not disappear. If only a single term is given there cannot be any question of a relation. The sole one I could think of is *identity* which is often considered a relation which an object has to itself. But identity in this sense is in my opinion no relation at all. It is but another name for object—the *same* object. To be sure, we can speak of two occurrences or two appearances of one and the same object, as well as we can speak of two objects being identical in certain respects, *e.g.* when they have the *same* property or are having the *same* relation to something. That, however, does not mean that there are two properties or relations that are identical, but solely that the *same* property belongs to both objects, or that several objects are having the *same* mutual relation as some other objects have—a property or a relation being something

that cannot be localized in space or time. Two different spots of colour may have the *same* shade of colour, we say. This, however, does not mean that there are *two* identical shades of colour, but that the same (one and the same) shade of colour is to be found in the two spots of colour. There is no relation of identity between two shades of colour, but the two different *spots* of colour have a property (*viz.* the shade of colour) *in common*. The *property* is the *same*—and *this* is the sense of the phrase " the two spots of colour are identical with respect to colour ". Two different objects may have a property in common ; but this fact does not constitute a relation of identity between the *objects*. As they are *two* they cannot be identical. Only the *colour* is identical in the two objects, *i.e.* they have the *same* colour, but there are not two colours with a relation of identity between them.

Further, it may be said that two different names have identical meanings, if they denote the *same* object. But it has presumably no sense to say, as it is often done, that an object is identical with itself, if " identical " should here mean (indicate) a relation. The expression " every object is identical with itself " can in my opinion solely mean, that no object is different from itself *qua* object, even if it may have various appearances in different situations. Either an object is the *same* object at various points of time or space, and if so it cannot have any relation to itself, or the object is not the same at various points of time or space, and if so there are several objects which may stand in various relations to each other, but none of which can have a relation of identity to any of the others.

If I am right in maintaining that no object or term can have any relation to itself, then all talk of so-called reflexive relations is senseless—a thesis that I would illustrate a little further by means of a few examples.

In the sentence " The blackboard is black " the word " blackboard " designates a blackboard. But in the sentence " The words ' the blackboard ' consists of 13 letters " the words " the blackboard " designate nothing, but are the very object about which something is predicated. The last sentence really degenerates into a predication by which a predicate, *viz.* " consists of 13 letters ", is predicated about the present object, *viz.* the word-*image* " the blackboard "—exactly as when somebody points at a blackboard and says " is black " or solely " black ". In the sentence " The words ' the blackboard ' consists of 13 letters " the words " the blackboard " is neither the subject-term nor the designation of the logical subject, but the very

subject itself, *i.e.* the object about which something is predicated. In this case the object, which is the *grammatical* subject of the statement, is exceptionally a word-*image*, but it is not a word, if "word" is to be understood as something that designates something.

Take then the sentence "The word 'long' is itself short". Here the word "long" designates nothing—although it is generally used to designate a *property*, which does not belong to the word-*image* "long". But in the sentence "The word 'short' is short" the second "short" designates a property that belongs to the word-image "short" itself. This, however, does *not* mean, that the word "short" here refers to itself. No, it designates a *property* that exceptionally belongs to the expression "short" itself, *i.e.* a property that belongs, not to the *word* "short", but to the *word-image* "short". And this word-image is not a word, *i.e.* it is not functioning as a symbol in a symbolizing process, but is simply an object that happens to have the property called short.

We are now able to dispose of and explain Grelling's (or Weil's) paradox according to which it is apparently possible to prove that a heterological word is not heterological. A property word (*e.g.* an adjective) is called heterological, if designating a property which does not belong to the word, *i.e.* the word-image, itself. The adjective "long" *e.g.* is a heterological word. If then we ask whether the word "heterological" itself is heterological, we seem to be led to contradictory answers: if it is heterological, it is not heterological; and if it is not heterological, then it is heterological. This paradox is in my opinion solved and explained when we realize that it is absurd to ask, whether the word "heterological" itself is heterological or not. The word "heterological" designates a property belonging (or not belonging) to a *word*, *viz.* a word that designates a property which belongs to the corresponding *word-image*. In order to decide whether a word is heterological or not, it is necessary to look at its word-image. But it has no sense to predicate the property "heterological" to the word-image "heterological", because the word "heterological" does not designate a property of a word-image, but a property of a word, namely a word designating a property. The paradox arises by a false identification of a word with its corresponding word-image—only the word can be meaningfully said to be heterological, whereas adjectives as "long" or "short" can only be ascribed to word-images. Words can neither be long or short, and word-images can neither be heterological nor homological.

A paradox of a similar character is Berry's paradox which runs as follows: In English, the lowest integer not nameable in fewer than 19 syllables is found to be nameable in 18 syllables. This seems to be a contradiction. The number is 111777, *i.e.* "one hundred and eleven thousand seven hundred and seventy seven". This number designation contains 19 syllables. But the number mentioned can also be unambiguously termed: "The least integer not nameable in fewer than nineteen syllables" where we have only 18 syllables.

The paradox seems to depend on the simple fact, that the relation between an object and its designation is assumed to be a one-one relation, whereas it is really a one-many relation. An object can be designated unambiguously by various different designations, and these need not contain the same number of syllables. There is no contradiction in saying: "The least integer whose number-name contains at least 19 syllables can also be named by another name that contains but 18 syllables".

Confer *e.g.* the description "the largest English town whose name contains at most two syllables" on the one hand, and "London" on the other. Here the town London is designated by the disyllable "London" as well as by the much longer description "the largest English town whose name contains at most two syllables", but there is no contradiction in naming this town in both ways. If, however, we assume that the relation of designation is one-one, then we have to face the paradox: "The largest English town whose name contains at most two syllables has a name that contains many more than two syllables". Really this statement concerns two different names, and there is no paradox at all. The appearance of paradox depends upon the wrong assumption that the town has but one name.

In a similar way it is, in my opinion, possible to dispose of and explain the paradox of "the relation which holds between R and S whenever R does not have the relation R to S"—my point being that it is meaningless to assume that a relation R can be its own relatum. But there are other paradoxes which depend on a wrong use of the word or concept of "all", and the solution of which seems to me to claim a kind of distinction of types—although a very simple one. The most famous of these paradoxes is, of course, Russell's paradox to which I will now turn.

It is well-known that this paradox concerns the class of classes which are not members of themselves. Apparently this class is a member of itself, if it is not a member of itself; and it is not a member of itself, if it is a member of itself. The solution that

Russell offers depends on the assumption that " a proposition about a class is always to be reduced to a statement about a function which defines the class, *i.e.* about a function which is satisfied by the members of the class and by no other arguments " (*Prin. Math.*, Vol. I, p. 62-63. 2nd edn.). Hence a class cannot, by the vicious-circle principle, significantly be an argument to its defining function, and therefore neither satisfies nor does not satisfy its defining function, and so is neither a member of itself nor not a member of itself.

While agreeing with his conclusion I am not at one with Russell in respect of the argument that leads to it. To me it seems simpler to resort to the following solution which at the same time gives an explanation of the appearance of the paradox.

A class is a collection of objects that are said to be members of the class. A class, therefore, must contain at least two members in order to be a class. Incidentally I may remark that for the same reason I consider expressions as " null-class " and " unit-class " to be merely ways of speaking (façons de parler) and not existing entities. If we put the particle " all " before a class name in a proposition about a class, we are not predicating something about the class as such but about its membership distributively, *i.e.* we are ascribing the predicate to each member of the class. The proposition " All S are P " says that each S severally has the property P—not that the class S has this property. " All horses are four-legged " means *e.g.* that each horse has four legs—not that the class of horses has four legs. This is the simple reason why a class is typically different from its members, and why a class cannot be a member of itself. Whatever other properties a class may have, it must contain at least two members in order to be a class at all. And a class cannot, no more than any other object, have any relation to itself, *e.g.* be a member of itself. That would mean that there could be a class without members, but that is absurd. Therefore, the expression " the class of classes which are not members of themselves " is a meaningless expression as well as the expression " the class of classes which are members of themselves ". And therefore the paradox does not arise at all.

To be more explicit : A class, S, presupposes that there are objects which are members of it. If these objects have a common property, P, then the class S can be defined as the collection of all the objects that have this property. But this property P cannot be a property of the *class* S because that would mean, that S should be one of the objects belonging to the class S,

since this class was defined as the collection of *all* the objects having the property P. If we try to escape this conclusion by defining: "The class S consists of all objects having the property P *except* the class S of these objects—plus this class" then we do not know what S means and therefore do not know what we should except from this class and then add to the other objects in the class in order to get S. This way of escape is therefore impossible. To be sure, we may try to define the class S "intensionally" without knowing whether there are members in it or not according to the scheme: "The class S consists of all objects that have the property P". But such definition is to my mind analytic or verbal. It solely means that *if* there are objects having the property P we will call them members of the class S. By this we have defined the use of the word "S" but not a class, because we do not know whether such class exists or not, *i.e.* whether there are objects having the property P or not, and if there are no such objects there is no class to be defined. The intensional definition, therefore, is not a definition of a *class* but solely of a *word*, a hypothetical class name. To sum up: We must distinguish between classes and class names. Classes must at least have two members, and can never be members of themselves. So-called intensional definitions of classes are really but definitions of class names and concerns solely our language, but not real collections of objects.

At this point it will, however, perhaps be expedient to remark that the word "all" and "class" have what Russell calls "systemic ambiguity". As already mentioned the proposition "All S are P" means that every single S has P (or belongs to the class P). The collection or class of P's may be said to be an entity of another logical type than the members of it. This type may be called type I. If several classes of type I are defined, we may collect them in a super-class and say that all classes of type I are members of the super-class which is, however, of another logical type that may be called type II. And so on, in accordance with Russell's simple theory of types.

The objects from which we start the construction of this hierarchy of classes may be of various kinds. They may, *e.g.* be individual things, or single properties of individual things, or relations between individual things. But if we speak of "a property of a property" or "a relation between relations" we are moving up in a hierarchy of properties or a hierarchy of relations—each new type being characterized by its members having members of the next lower type. The word "all" also changes its designative meaning or function as we ascend the

hierarchy, the word "all" placed before a class-name in a proposition referring always to the *members* of the class named and never to that class itself.

I think there is no need of any more complicated theory of logical types. But what about the so-called reflexibility of languages?

As is well-known we can, in a certain sense, speak *of* a language *in* the language itself, and Carnap has even, in his "Logical Syntax of Language", taken the trouble to construct an artificial language containing its own syntax. May such languages then not be reflexive phenomena?

I don't think so, and shall now try to show why. To do so it will, I believe, be sufficient to consider our everyday language. This is, as Tarski somewhere remarks, an "open" language in the sense that its vocabulary as well as its syntax may be changed or extended as the language develops. But it is in my opinion "open" in an even more fundamental sense: If I pronounce a sentence S_2 about another sentence S_1, then S_2 is a new linguistic phenomenon—even if it is perhaps similar to S_1 in such a way that there is but a numerical difference between S_1 and S_2. If S_2 consists of the "same" words as S_1, and if these words are arranged in the "same" order as these words are arranged in S_1, then S_2 may be said to have spoken about S_1 without changing either the vocabulary or the syntax of the language in which the two sentences are spoken. But the vocabulary and the syntax of a language is but an abstraction, a structure, while the language itself is a concrete phenomenon in time or space—or more strictly: an unlimited course or lapse of concrete phenomena. As long as a language is used, it is "open", even if its new sentences are built of elements of one and the same vocabulary, and according to one and the same syntax. Therefore a sentence can never say anything about itself. It is, indeed, not a sentence before it is finished, and if something is to be said about it, then this something must be said after the sentence is finished. The case is similar to Royce's map of England which he thought should contain a map of the map, and a map of the map of the map, and so on, in order to be complete. That is not the case. There should not be any map of a map before the first map is designed. And when the first map is designed there should only be a map of this map. In order to be complete a map of England shall never contain a (complete) map of itself. An infinite number of maps of maps of England never exists, and therefore should never be mapped, but any number of really designed maps of maps *can* be mapped,

and the latest of them is always the correct and complete map corresponding to the prevailing state of affairs—until it is finished itself. Then a new map should be made containing the preceding one. And when that one is finished a new one, and so on. As long as England and the mapping of England is not finished new maps ought to be made, but the infinity is always a progressive one and never contains a finished infinity. In a similar way the language is always progressing. Any repetition of a sentence is a new occurrence of the sentence, and a sentence cannot say anything of a *sentence* that is not yet finished, but at most of the abstract *form* which the unfinished sentence may be assumed to have—when it is finished. In the same way a language cannot say anything about this language as a whole, because the pronouncement about the "whole" language will add something to it. But one part of a language may very well say something about another (concrete or abstract) part of the language, provided that this other part already exists. And the parts may very well be similar, but can never be identical. This is but another way to say that a sentence can never refer to itself, never be its own subject.

Similarly, self-consciousness is never a reflexive phenomenon. The act of consciousness in which we are conscious of ourselves being conscious of something is always a *new* act of consciousness, and this new act of consciousness can only be conscious in yet another new act of consciousness. Therefore an act of consciousness can never be conscious of itself. But we, as persons existing in time can successively be conscious of our preceding acts of consciousness, although we can never be conscious of our present act of consciousness before it is completed. That would mean to jump over our own shadow, or rather: that would mean to be conscious of something without there being anything to be conscious of. "Introspection is always retrospection", as Professor Ryle has said, I think.

The main point is, in my opinion, that knowing is a temporal process, and that any act of knowing must exist before we can know it—otherwise we could know or speak about an act of knowing that does not yet exist in the sense that it would be nothing at all.

In the same way we can, I believe, dispose of other seemingly reflexive phenomena as, *e.g.* "a theory of theories", "a concept of concepts", "knowledge of knowledge", and so on. But what about the so-called reflexive classes of numbers? Are they not genuinely reflexive? I don't think so, and I believe the so-called paradoxes of the infinite series of numbers can be

shown to depend on a failure to distinguish between two simple forms of natural numbers. But this distinction being a bit more complicated than the above mentioned I will leave it for another occasion. What I have intended here is merely to show the possibility of eliminating some of the paradoxes which many logicians nowadays seem to have accustomed themselves to in such a degree that they almost consider such anomalies as established matters of course. As far as my memory serves me, Lord Russell somewhere says that " paradoxes are the experiments of logic ". I don't agree. To my mind paradoxes are rather " traps of logic ", and I don't like to see logicians trapped—not even in their own traps. I, therefore, have tried to destroy some of these traps. Whether I have succeeded or not I leave to you to decide. But should my present paper merely serve as a warning against the traps, and as a reminder of being cautious, I should consider even such modest effect as not being quite unsatisfactory.

University of Copenhagen

R. M. MARTIN

ON NON-TRANSLATIONAL SEMANTICS

Following Carnap and Church, one can distinguish between a *formalized logistic system* (or calculus) and a *formalized language-system* (or interpreted language) somewhat roughly as follows [1]. The former is determined by rules which refer exclusively to symbols and expressions, regarded in abstraction from any specific interpretation. A language-system, on the other hand, is a logistic system with a fixed, determinate interpretation or assignment of denotata given to certain of its expressions. Amongst the totality of rules constitutive of a language-system one can distinguish between those which are *syntactical* and those which are *semantical* in character. Ordinarily, the syntactical rules refer only to the expressions, whereas semantical rules are concerned with the interrelations of the expressions with denotata. A logistic system is wholly determined by syntactical rules. It is therefore sometimes also called a *syntactical system,* and the language in which it is formulated is often called a *syntactical meta-language.* Language-systems, on the other hand, are sometimes called *semantical systems,* being determined by syntactical and semantical rules together. The meta-language in which a semantical system is formulated is often called a *semantical meta-language.*

Roughly speaking, any formalized system, whether syntactical or semantical, whether an object- or a meta-language, consists of the following:

(a) a complete specification of the *primitive vocabulary,*

(b) an explicit definition (recursive or otherwise) of what it means to be a *formula,* and possibly *term,* of that system,

(c) a finite or denumerable list of formulae as *axioms* or *primitive sentences,*

(d) some statements about the language, the *rules of inference,* telling us the circumstances under which a formula or sentence is to be regarded as provable from or an immediate consequence of a formula or formulae,

---

[1] See R. Carnap, *Introduction to Semantics* (Harvard University Press, Cambridge, Mass., 1942), esp. pp. 22-29; and A. Church, "The need for abstract entities in semantic analysis", *Proceedings of the American Academy of Arts and Sciences* 80 (1951), p. 100.

(e) a list of formulae or sentences explicitly shown to be *theorems*, i.e., shown to be provable from the axioms by means of a finite number of applications of the rules of inference,

(f) a list of statements about the language allowing us to *define* or abbreviate expressions, especially long ones, in specified ways,

and, in the case of semantical systems,

(g) a list of statements about the language which assign explicitly denotata to certain of the constituent expressions or which otherwise tell us how various expressions are semantically interrelated.

This rough description is of course not intended as an exact definition. But any exact definition or characterization of a formalized system (of an appropriate kind) must be such as to accord with this rough description in essential respects.

All the language-systems discussed in this paper, either as object- or as syntactical or semantical meta-languages, are of the kind known as simple, applied, logical languages of first order with or without identity [2]). They thus all contain certain basic logical ingredients, such words or phrases as 'and', 'or', 'if... then...', 'for all *a*', etc. These words are to have the meanings associated with them in the usual classical, two-valued logic of truth-functions and quantifiers. The logical axioms and rules presupposed in the languages discussed here are thus of the familiar kind.

A syntactical meta-language is of course a language and may, if desired, be formulated so as to presuppose an extremely simple basic logic [3]). As primitives one needs only names of the various basic symbols of the object-language, a functor (or functional symbol) expressing concatenation (the operation of forming the expression *a* followed by *b* from the constituent expressions *a* and *b*), and the expressions for the logical notions including identity. The variables of this kind of a syntactical meta-language range merely over the expressions of the object-language, these expressions constituting a denumerable fundamental domain of entities each of finite length.

The structure of semantical meta-languages, however, is in general more complicated [4]). Semantics presupposes syntax and contains it as a part,

---

[2]) See A. Church, *Introduction to Mathematical Logic (Princeton Universtity* Press, Princeton, N.J., 1944), esp. p. 37.

[3]) See L. Chwistek, *The Limits of Science* (Kegan Paul, London, 1948), pp. 83 - 100 and 162 - 191; and W. V. Quine, *Mathematical Logic* (New York, Norton, 1940), pp. 291-305.

[4]) See especially A. Tarski, "Der Wahrheitsbegriff in den formalisierten Sprachen," *Studia Philosophica* 1 (1936), pp. 261-405.

and a semantical meta-language therefore contains a syntactical one as a part. But semantical meta-languages contain a good deal more also. The semantical meta-languages ordinarily considered contain, roughly speaking, four parts, a logical part, the syntactical part, a translation part, and a semantical part. The logical and syntactical parts are essentially the same as in a syntactical meta-language of the kind described above, although the basic logic must be slightly broader. Also the semantical meta-languages contain a translation *in toto* of their object-languages. The translation part may be just the object-language itself or any language which corresponds with the object-language in a suitable way. Finally, within the semantical part of the meta-language, one interrelates the expressions of the object-language with their denotata in a systematic way.

All semantical meta-languages which have heretofore been studied in any detail seem to conform to this general pattern. In particular, they all contain a translation *in toto* of the object-language to which they are applied. We may speak of any such semantical meta-language as a *translational* meta-language, and the semantics involved as a translational semantics. All semantical meta-languages heretofore formulated seem to have been translational in character.

In this paper the foundations for a semantical theory will be given which, because it lacks this crucial feature, will be called a *non-translational* semantics; and the meta-language to be formulated will be called a non-translational, semantical meta-language. This meta-language will afford a radical departure from the kinds of semantical meta-languages studied by Tarski, Carnap, and others. In some respects, it is less powerful and certain principles of ordinary, translational semantics will have to be sacrificed. Bearing in mind the vital role these principles have played, it is perhaps somewhat remarkable that so much can be accomplished without them.

By a *name* let us mean any expression of the object-language which is either a one-place predicate constant or a one-place abstract. One name *a* can be said to *comprehend* a name *b* if and only if, informally speaking, *a* applies to every object to which *b* applies [5]). Thus, e.g., 'lover of peace' may be said to comprehend 'Quaker' in an appropriate formalism. The sole semantical primitive needed in the construction of non-translational semantics is the relation of comprehension. Several important semantical ideas are definable in terms of comprehension including a semantical truth-concept.

[5]) See the author's abstract, "A semantics without ontology," *Journal of Symbolic Logic* 17 (1952), pp. 157 - 158.

As object-language let us take then any first-order language *L*. This may be one of the languages recently studied by Carnap, Hempel, and others as a basis for investigations in the logic of science [6]). Or *L* may be taken as one of the famous mathematical languages, such as a language based upon the theory of types or upon a set theory of the kind studied by Zermelo or von Neumann [7]). For convenience let us presuppose that *L* contains no primitive individual constants or proper names and no primitive functors, but only some predicate- or relational constants. This assumption is inessential but helps to simplify the semantical formulations below.

We can presuppose that the syntactical meta-language of *L* is formulated in the narrow way suggested above. Let us use '*a*', '*b*', etc., as variables of the syntactical meta-language ranging over the expressions *L*. In addition to these variables, the syntactical meta-language contains names of each of the primitive symbols of *L*. These names may also be called the *structural descriptions* respectively of the symbols which they name. Ordinarily *L* may be supposed to contain only a very few symbols primitively, perhaps only half a dozen or so. In addition to the structural-descriptive names, the syntactical meta-language contains the identity sign and a symbol for the operation of concatenation. The axioms of this elementary syntax are certain very simple statements giving the basic properties of concatenation [8]).

The non-translational semantical meta-language for *L* consists of elementary concatenation theory augmented by the theory of comprehension. The total vocabulary of the semantical meta-language thus consists merely of that of the syntax language together with a symbol 'Cmprh' for the relation of comprehension.

Within the elementary syntax of *L* we can define such notions as 'theorem of *L*', 'sentential function of one variable of *L*', 'predicate constant of *L*', etc. With the addition of 'Cmprh' several important semantical notions become definable. Thus, an expression *a* of *L* can be said to be a *name* if and only if there is at least one *b* such that *a* Cmprh *b* or *b* Cmprh *a*. Expressions *a* and *b* can be said to be *coextensive* with each other if and only if they mutually comprehend each other. An expression *a* can be said

---

[6]) See R. Carnap, *The Logical Foundations of Probability*, (University of Chicago Press, 1950), pp. 58 - 68; and C. G. Hempel and P. Oppenheim, "Studies in the logic of explanation," *Philosophy of Science* 15 (1948), pp. 135 - 175.

[7]) E. Zermelo, "Untersuchungen über die Grundlagen der Mengenlehre," *Mathematische Annalen* 65 (1908), pp. 107 - 128, and J. von Neumann, "Die Axiomatisierung der Mengenlehre," *Mathematische Zeitschrift* 27 (1928), pp. 669 - 752.

[8]) See Tarski, *loc. cit.*, p. 289. An additional assumption is also required.

to be *semantically null* if and only if *a* is comprehended by every name, and *semantically universal* if and only if it comprehends every name. An expression can be called a *unit name* if and only if it is a non-null name and is comprehended by every non-null name which it comprehends. An expression *a* can be said to be a *semantical sum* of expressions *b* and *c* provided *a* Cmprh *b*, *a* Cmprh *c*, and for every *d*, if *d* Cmprh *b* and *d* Cmprh *c* then *d* Cmrph *a*. In a somewhat similar way one can define the notions of being a semantical product of two expressions and of being a semantical negative of an expression.

The general theory of comprehension contains a kind of semantical Boolean algebra, in which the predicate constants or one-place abstracts of *L* are the basic elements. The truth-concept for *L* is readily definable within such a theory. Thus, an expression *a* can be said to be *true in L* if and only if *a* is a sentence of *L* and there exists a variable *b* such that the abstract consisting of *b* followed by '$\in$', followed by *a* is a universal name [9]).

One of the most important features of non-translational semantics is that the requirement of *adequacy* (in essentially the sense of Kotarbiński and Tarski) for the truth-definition must be abandoned. The reason is that it cannot even be stated within the non-translational semantical formalism, because this contains only structural descriptions, not translations, of the expressions of *L*. To state the requirement of adequacy demands that within the semantical meta-language one have both, i.e., that one can both mention and use (in effect) the expressions of *L*. Thus, in a certain sense, the adequacy of the truth-concept defined above must be abandoned. It is to be noted, however, that the truth-definition there is *in accord with* the requirement of adequacy. In other words, the meaning given to the semantical primitive 'Cmprh' is such that the resulting truth-concept is essentially the same as one within a semantical formalism for which adequacy could be stated and proved. Thus, it is not the case that the requirement of adequacy is in any way denied here or its negative assumed. That the truth-concept defined is clearly in accord with the requirement of adequacy, however, cannot be explicitly proved; this must rather rest on intuitive grounds concerning the meaning of 'Cmprh'.

The rules of non-translational semantics for a given *L* may be described very roughly as follows. They include several rules providing for the underlying Boolean algebra of comprehension, a rule to the effect that every

[9]) The inverted epsilon is Peano's symbol for class abstraction. It is presupposed here as a primitive of *L*. The rule of *L* governing it is an adaptation of D4.1 of the author's "A homogeneous system for formal logic," *Journal of Symbolic Logic* 8 (1943), pp. 1 - 23.

name is a predicate constant and conversely, a suitable existence rule, and some rules providing for the semantical properties of abstracts. For each non-logical axiom of L a corresponding assumption is required in the semantical meta-language stating that such and such an abstract has such and such a semantical property. E.g., if L is taken as the formalized Zermelo–Skolem set theory, corresponding to the axiom concerning the existence of the null set, we could assume within the semantical meta-language that the abstract

$$`x `\exists' (y) \sim y \varepsilon x'$$

is non-null. The task of non-translational semantics, however, is not to tell us specifically which sentences of the object-language are true, but only to provide the general framework for a semantical truth-definition. The rules of non-translational semantics contain nothing dubious and should in no way depend upon the non-logical axioms of a particular object-language. Hence we shall prefer to take as additional hypotheses wherever needed the assumptions corresponding to the non-logical axioms of the object-language. The use of such hypotheses wherever needed is akin to that of *Principia Mathematica* with respect to the Axiom of Infinity and the Multiplicative Axiom. On such assumptions the basic semantical theorems concerning L are readily forthcoming. In particular, on such assumptions one can prove the consistency of the object-language L by an adaptation of the familiar methods of translational semantics.

Finally, on the supposition that the axioms of the underlying syntax are consistent, one can readily show that the rules of non-translational semantics are also consistent. This relative consistency proof assures us of a secure first-order semantics applying to any first-order object-language of any complexity. And in particular, the non-translational semantical meta-language can be formulated in such a way as to provide a consistent semantics for itself.

Considerable philosophical interest attaches to a semantics of this kind. For one thing, it seems to afford a very simple and economical formulation of denotational semantics. With only slight changes it appears to give also a solution to the problem of gaining a nominalistic truth-definition in the strict sense of Goodman and Quine [10]). Also it provides a denumerable semantics irrespective of whether the object-language L is denumerable or

[10]) See N. Goodman and W. V. Quine, "Steps toward a constructive nominalism," *Journal of Symbolic Logic* 12 (1947), pp. 105 - 122. Also N. Goodman, *The Structure of Appearance* (Harvard University Press, Cambridge, Mass., 1951), esp. pp. 31 - 41.

not. Thus in a certain sense non-translational semantics may be said to avoid the situation in the foundations of mathematics known as the paradox of Skolem, whereas all semantical methods heretofore, for denumerable or non-denumerable L, have been subject to it fundamentally [11]).

University of Pennsylvania
PHILADELPHIA 4, Penna
U.S.A.

---

[11]) See Th. Skolem, "Einige Bemerkungen zur axiomatischen Begründung der Mengenlehre," *Wissenschaftliche Vorträge gehalten auf dem Fünften Kongress der Skandinavischen Mathematiker in Helsingfors vom 4. bis 7. Juli 1922* (Helsingfors, 1923), pp. 217 - 232.

THE JOURNAL OF SYMBOLIC LOGIC
Volume 22, Number 1, March 1957

LANGUAGES IN WHICH SELF REFERENCE IS POSSIBLE[1]

RAYMOND M. SMULLYAN

1. Introduction. This paper treats of semantical systems $\boldsymbol{S}$ of sufficient strength so that for any set W definable in $\boldsymbol{S}$ (in a sense which will be made precise), there must exist a sentence X which is true in $\boldsymbol{S}$ if and only if it is an element of W.[2] We call such an X a *Tarski* sentence for W. It is the sentence which (in a purely extensional sense) says of itself that it is in W.[3] If W is the set of all expressions not provable in some syntactical system C, then X is the Gödel sentence which is true (in $\boldsymbol{S}$) if and only if it is not provable (in C). We provide a novel method for the construction of these sentences, which yields sentences particularly simple in structure. The method is applicable to a variety of systems, including a form of elementary arithmetic, and some systems of protosyntax self applied.[4] In application to the former, we obtain an extremely simple and direct proof of a theorem, which is essentially Tarski's theorem that the truth set of elementary arithmetic is not arithmetically definable.

The crux of our method is in the use of a certain function, the 'norm' function, which replaces the classical use of the *diagonal* function. To give a heuristic idea of the norm function, let us define the norm of an expression E (of informal English) as E followed by its own quotation. Now, given a set W (of expressions), to construct a sentence X which says of itself that it is in W, we do so as follows:

W contains the norm of 'W contains the norm of.' This sentence X says that the norm of the expression 'W contains the norm of' is in W. However,

Received May 3, 1956.

[1] I wish to express my deepest thanks to Professor Rudolph Carnap, of the University of California at Los Angeles, and to Professor John Kemeny and Dr. Edward J. Cogan, of Dartmouth College, for some valuable suggestions. I also wish to thank the referee for some very helpful revisions.

[2] By a semantical system $\boldsymbol{S}$, we mean a set E of *expressions* (strings of signs), together with a subset S of expressions called *sentences* of $\boldsymbol{S}$ (determined by a set of rules of formation), together with a subset T of S, of elements called *true* sentences of $\boldsymbol{S}$ (determined by a set of 'rules of truth' for $\boldsymbol{S}$).

[3] (At the referee's suggestion) — In an extensional system, the only way we can translate the meta-linguistic phrase '$\boldsymbol{X}$ says that $\boldsymbol{X} \in W$' is by the phrase '$\boldsymbol{X}$ is true if and only if $\boldsymbol{X} \in W$'. Thus the requirement for $\boldsymbol{X}$ to be a Tarski sentence for W, is exceedingly weak; any sentence $\boldsymbol{X}$ which is either both true and in W, or false and not in W, will serve. However, this is as much as we need of a Tarski sentence (for undecidability results). If we were considering an *intensional* system, then we would define a Tarski sentence for W as a sentence $\boldsymbol{X}$ which not only is true if and only if $\boldsymbol{X} \in W$, but which actually expresses the proposition that $\boldsymbol{X} \in W$.

[4] The former will be carried out in this paper and the latter in a forthcoming paper, *Systems of protosyntax self applied.*

the norm of this expression is X itself. Hence X is true if and only if $X \in W$.[5]

This construction is much like one due to Quine.[6] We carry it out for some formalized languages. In section 2, which is essentially expository, we construct a very precise, though quite trivial semantical system $\boldsymbol{S}_P$, which takes quotation and the norm function as primitive. The study of this system will have a good deal of heuristic value, inasmuch as $\boldsymbol{S}_P$, despite its triviality, embodies the crucial ideas behind undecidability results for deeper non-trivial systems. We then consider, in section 3, the general use of the norm function, and we finally apply the results, in section 4, to a system $\boldsymbol{S}_A$, which is a formal variant of elementary arithmetic. This variant consists of taking the lower functional calculus with class abstractors, rather than quantifiers, as primitive. This alteration, though in no way affecting the strength of the system, nevertheless makes possible the particularly simple proof of Tarski's or Gödel's theorem, since the arithmetization of substitution can thereby be circumvented quite simply.

By the norm of an expression E (of $\boldsymbol{S}_A$) we mean E followed by its own Gödel *numeral* (i.e., the numeral designating its Gödel number). Now, given any set W of expressions whose set of Gödel numbers is arithmetically definable, we show quite easily the existence of an expression H of class abstraction, such that for any expression E, H followed by the Gödel *numeral* of E is a true sentence if and only if the *norm* of E is in W. Then, if we follow H by its own Gödel numeral h, the resulting sentence Hh (which is the norm of H) is true if and only if it is in W. This is a rough sketch of our procedure.

2. The preliminary system $\boldsymbol{S}_0$ and the semantical system $\boldsymbol{S}_P$. In this section, we formalize the ideas behind the preceding heuristic account of the norm function. For convenience, we first construct a preliminary system $\boldsymbol{S}_0$, whose expressions will be built from the three signs 'φ', '*', and 'N'. The second sign will serve as our formal quotation mark, since we reserve ordinary quotation marks for meta-linguistic use. The sign 'N' will be endowed with the same meaning as 'the norm of.' The sign 'φ' will be an undefined predicate constant. For any property (set) P of expressions of $\boldsymbol{S}_0$, we then define the semantical system $\boldsymbol{S}_P$ by giving a rule of truth for $\boldsymbol{S}_P$. For any P, 'φ' will be interpreted in $\boldsymbol{S}_P$ as designating P.

---

[5] In contrast with this construction, let us define the *diagonalization* of E as the result of substituting the quotation of E for all occurrences of the variable 'x' in E. Then the following Tarski sentence for W (when formalized) is the classical construction: W contains the diagonalization of 'W contains the diagonalization of x'. This latter construction involves *substitution* (inherent in diagonalization), whereas the norm function involves *concatenation* (the norm of E being E *followed by* its quotation), which is far easier to formalize (cf. Section 5).

[6] 'Yields falsehood when appended to its quotation' yields falsehood when appended to its quotation. This is Quine's version of the famous semantical paradox.

Signs of $\boldsymbol{S}_0$: φ, *, N.
Preliminary definitions: (1) By an *expression* (of $\boldsymbol{S}_0$) we mean any string built from the three signs of $\boldsymbol{S}_0$. (2) By the (formal) quotation of an expression, we mean the expression surrounded by stars. (3) By the *norm* of an expression, we mean the expression followed by its own (formal) quotation.
Formation rules for (individual) designators
(1) The quotation of any expression is a designator.
(2) If E is a designator, so is ⌜NE⌝ (i.e., 'N' followed by E).
Alternative definition
(1)′ A designator is an expression which is either a quotation (of some other expression) or a quotation preceded by one or more 'N's.
Rules of designation in $\boldsymbol{S}_0$
R1. The quotation of an expression E designates E.
R2. If E_1 designates E_2, then ⌜NE_1⌝ designates the *norm* of E_2.
Definition of a sentence of $\boldsymbol{S}_0$
(1) A sentence of $\boldsymbol{S}_0$ is an expression consisting of 'φ' followed by a designator.
The semantical system $\boldsymbol{S}_P$
For any property P, we define the semantical system $\boldsymbol{S}_P$ as follows:
(1) The rules for designators, designation and sentence formation in $\boldsymbol{S}_P$ are the same as in $\boldsymbol{S}_0$.
(2) The rule of truth for $\boldsymbol{S}_P$ is the following:
R3. For any designator E, ⌜φE⌝ is true in $\boldsymbol{S}_P \underset{\text{df}}{=}$ the expression designated by E (in $\boldsymbol{S}_P$) has the property P.

THEOREM 2.1. There exists an expression of $\boldsymbol{S}_0$, which designates itself.

PROOF. '*N*' designates 'N' [By Rule 1].
Hence 'N*N*' designates the norm of 'N' [By Rule 2] which is 'N*N*'.
Thus 'N*N*' designates itself.

THEOREM 2.2. There exists a sentence G of $\boldsymbol{S}_0$, such that for any property P, G is true in $\boldsymbol{S}_P \iff G$ has the property P.

PROOF. 'N*φN*' designates 'φN*φN*' [By R1 and R2].
Thus G, viz., 'φN*φN*' is our desired sentence.

REMARK. G is, of course, the formalized version of 'W contains the norm of 'W contains the norm of'.' 'φ' is but an abbreviation of 'W contains,' and 'N' abbreviates 'the norm of.'[7]

---

[7] If we wished to construct a miniature system $\boldsymbol{L}_P$ which formalizes the diagonal function in the same way as $\boldsymbol{S}_P$ does the norm function, we take four signs, viz., 'φ', '*', 'D', 'x', and the rules R_1, (same as $\boldsymbol{S}_P$), R_2: If E_1 designates E_2, then ⌜DE_1⌝ designates the diagonalization of E_2 (i.e., the result of replacing each occurrence of 'x' in E_2 by the quotation of E_2), R_3: If E_1 designates E_2, then ⌜φE_1⌝ is a sentence and is true in $\boldsymbol{L}_P$ if and only if E_2 has the property P. Then the expression of Theorem 2.1, which designates itself, is 'D*Dx*', and the Tarski sentence (of Theorem 2.2) for P is 'φD*φDx*'.

COROLLARY 2.3. P cannot be coextensive with the set of all false (non-true) sentences of $\boldsymbol{S}_P$, nor is P coextensive with the set of all expressions of $\boldsymbol{S}_P$ which are not true sentences of $\boldsymbol{S}_P$.

2.4. A STRONGER FORM OF THEOREM 2.2. By a *predicate* we mean either 'φ' or 'φ' followed by one or more 'N's'.

We say that an expression E *satisfies* a predicate H (in $\boldsymbol{S}_P$) if H followed by the quotation $\ulcorner$*E*$\urcorner$ of E, is true in $\boldsymbol{S}_P$. Lastly, we say that a set W of expressions of $\boldsymbol{S}_0$ is *definable* (in $\boldsymbol{S}_P$) if there exists a predicate H which is satisfied by all and only those expressions which are in W.

It is worth noting at this point, that if E_1 designates E, then E satisfies H if and only if $\ulcorner HE_1\urcorner$ is true. This follows from R3 by induction on the number of N's occuring in H.

For any set W, we let $\eta(W) \underset{\text{df}}{=}$ set of all expressions whose norm is in W.

LEMMA 2.5. If W is definable in $\boldsymbol{S}_P$, then so is $\eta(W)$.

PROOF. Let H be the predicate which defines W (i.e., which is satisfied by just those elements which are in W). Then H followed by 'N' will be satisfied by precisely those elements which are in $\eta(W)$. Thus $\eta(W)$ is definable (in $\boldsymbol{S}_P$).

We can now state the following theorem, of which Theorem 2.2 is a special case.

THEOREM 2.6. For *any* set W definable in $\boldsymbol{S}_P$, there is a sentence X which is true in $\boldsymbol{S}_P$ if and only if $X \in W$.

PROOF. Assume W is definable. Then so is $\eta(W)$ [by Lemma]. Hence there exists a predicate H such that for any expression E, $\ulcorner$H*E*$\urcorner$ is true (in $\boldsymbol{S}_P$) $\iff$ E$\in\eta(W)$

$$\iff \ulcorner \text{E*E*}\urcorner \in W.$$

Taking E = H, $\ulcorner$H*H*$\urcorner$ is true $\iff \ulcorner$H*H*$\urcorner \in W$.

Thus X, viz., $\ulcorner$H*H*$\urcorner$, is our desired sentence.

REMARK. Theorem 2.6 says (in view of the truth functionality of the biconditional) no more nor less than this: each set definable in $\boldsymbol{S}_P$ either contains some truths of $\boldsymbol{S}_P$ or lacks some falsehoods.

COROLLARY 2.7. The set of false sentences of $\boldsymbol{S}_P$ is not definable in $\boldsymbol{S}_P$, nor is the complement (relative to the set of all *expressions* of $\boldsymbol{S}_P$) of the set of true sentences of $\boldsymbol{S}_P$ definable in $\boldsymbol{S}_P$.

COROLLARY 2.8. Suppose we extend $\boldsymbol{S}_P$ to the enlarged semantical system $\boldsymbol{S}'_P$ by adding the new sign '$\sim$', and adding the following two rules:

R4. If X is a sentence, so is $\ulcorner\sim X\urcorner$.

R5. $\ulcorner\sim X\urcorner$ is true in $\boldsymbol{S}'_P \iff X$ is not true in $\boldsymbol{S}'_P$.

Then in this system $\boldsymbol{S}'_P$, the truth set of $\boldsymbol{S}'_P$ is not definable.

PROOF. For $\boldsymbol{S}'_P$ has the property that the complement of any set definable in $\boldsymbol{S}'_P$ is again definable in $\boldsymbol{S}'_P$, since if H defines W, then $\ulcorner\sim$H$\urcorner$ defines

the complement of W. Hence the truth set is not definable, since its complement is not definable by Corollary 2.7.

REMARK. S'_P is about as simple a system as can be constructed which has the interesting property that the truth set of the system is not definable within the system and that, moreover, any possible extension of S'_P will retain this feature. By an extension, we mean any system constructed from S'_P by possibly adding additional signs, and rules, but retaining the old rules in which, however, the word 'expression' is re-interpreted to mean an expression of the enlarged system. Likewise, if we take any extension of S_P, then, although we may greatly enlarge the collection of definable sets, none of them can possibly be co-extensive with the set of false sentences of the extension.

2.8. EXTENSION OF S_P TO A SEMANTICO-SYNTACTICAL SYSTEM S^C_P.

Suppose now that we select an arbitrary set of sentences of S_0 and call them *axioms*, and select a set of rules for inferring sentences from other sentences (or finite sets of sentences). The axioms, together with the rules of inference, form a so-called syntactical system, or calculus C. Let S^C_P be the ordered pair (S_P, C). Thus S^C_P is a mathematical system, or interpreted calculus. We let T be the set of true sentences of S_P (also called true sentences of S^C_P) and Th, the set of sentences provable in C (also called provable sentences, or theorems, of S^C_P). We already know that the complement $\overline{\text{T}}$ of T (relative to the set of expressions) is not semantically definable in S^C_P (i.e., not definable in S_P); however $\overline{\text{Th}}$ may well happen to be. If it is, however, then we have, as an immediate corollary of 2.6, the following miniature version of Gödel's theorem:

THEOREM 2.9. If the set $\overline{\text{Th}}$ is semantically definable in S^C_P, then either some sentence true in S^C_P cannot be proved in S^C_P or some false sentence can be proved.

This situation is sometimes described by saying that S^C_P is either semantically incomplete or semantically inconsistent.

2.10. We can easily construct a system S^C_P obeying the hypothesis of theorem 2.9 as follows: Before we choose a property P, we *first* construct a completely arbitrary calculus C. Then we simply define P to be the set of all expressions not provable in C. Then 'φ' itself will be the predicate which semantically defines $\overline{\text{Th}}$ in S^C_P, and the sentence G, viz., '$\varphi\text{N}^*\varphi\text{N}^*$', of Theorem 2.2 will be our Gödel sentence for S^C_P, which is true if and only if not provable in the system. In fact, for purposes of illustration, let us consider a calculus C with only a finite number of axioms, and no rules of inference. Thus the theorems of C are the axioms of C. Now, if G was included as one of the axioms, it is automatically false (in this system), whereas if G was left out, then it is true, by very virtue ofbeing left out. Thus, this system is, with dramatic clarity, obviously inconsistent or incomplete.

REMARK. Suppose that we take P to be the set of sentences which *are* provable in C. Then G becomes the Henkin sentence for the system $\boldsymbol{S}_P^C$, which is true in this system, if and only if G *is* provable in $\boldsymbol{S}_P^C$. Is G true in $\boldsymbol{S}_P^C$? This obviously depends on C. If, for example, we take C such that its set of axioms is null, then G is certainly both false and non-provable. An example of a choice of C (other than an obvious one in which G itself is an axiom) for which G is true is the following: We take for our single axiom A_1, the expression 'φ*φN*φN**'. We take a single rule R: If two designators E_1 and E_2 have the same designatum in $\boldsymbol{S}_0$, then $\ulcorner\varphi\mathrm{E}_2\urcorner$ is directly derivable from $\ulcorner\varphi\mathrm{E}_1\urcorner$. This rule is 'reasonable' in the sense that it does preserve truth in $\boldsymbol{S}_P$.

Now, is G, viz., 'φN*φN*', true in this $\boldsymbol{S}_P^C$ or not? It is true, providing it is provable. Now since 'N*φN*' and '*φN*φN**' both have the same designatum 'φN*φN*', then 'φN*φN*' is immediately derivable from 'φ*φN*φN**' by R, i.e., G is immediately derivable from A_1, hence G is provable, and hence also true.

We now consider whether or not this system $\boldsymbol{S}_P^C$ is semantically consistent. We have already observed that rule R does preserve truth, so the question reduces to whether or not A_1 is true. Well, by R3 of $\boldsymbol{S}_P$, A_1 is true precisely in case 'φN*φN*' has the property P, i.e., precisely in case G is provable, which it is.

3. Semantical Systems with predicates and individual constants. In this section, we consider any semantical system $\boldsymbol{S}$, of which certain expressions called *predicates* and certain expressions called (individual) *constants* are so related that any predicate followed by any constant is a sentence of $\boldsymbol{S}$. We also wish to have something of the nature of a *Gödel correspondence* g which will assign a unique constant $g(\mathrm{E})$ to each expression E so that $g(\mathrm{E})$ will be peculiar to E — i.e., we consider a 1—1 correspondence g whose domain is the set of all expressions, and whose range is a subset (proper or otherwise) of the individual constants. We shall often write '$\mathring{\mathrm{E}}$' for '$g(\mathrm{E})$'. By the *norm* of E, we mean $\mathrm{E}\mathring{\mathrm{E}}$ (i.e., E followed by $g(\mathrm{E})$).[8] [9] For a set W of expressions of $\boldsymbol{S}$, by $\eta(W)$ we mean the set of all E whose norm is in W.

---

[8] The word 'norm' was suggested by the following usage: In Mathematics, when we have a function f(x, y) of two arguments, the entity f(x, x) is sometimes referred to as the norm of x — e.g., in Algebra the norm of a vector ϵ is f(ϵ, ϵ), where f is the function which assigns to any pair of vectors the square root of their inner product. In this paper, the crucial function f (which allows us to introduce a notion of representability) is the function which assigns, to each pair (E_1, E_2) of expressions, the expression $\mathrm{E}_1\mathring{\mathrm{E}}_2$. Thus for this f, f(E, E) is our norm of E.

[9] Professor Quine has kindly suggested that I remark that the norm of a predicate is a sentence which says in effect that the predicate is *autological*, in the sense of Grelling (i.e., that the predicate applies to itself.)

A predicate H is said to define the set W (relative to g, understood) if W consists of all and only those expressions E such that HE is true. The following theorem, though quite simple, is basic.

THEOREM 3.1. For any set W of expressions of $\boldsymbol{S}$, a sufficient condition for the existence of a Tarski sentence for W is that $\eta(W)$ be definable.

PROOF. Suppose $\eta(W)$ is definable. Then for some predicate H and for any E $\mathrm{H\mathring{E}}$ is true $\iff \mathrm{E}\epsilon\eta(W)$

$$\iff \mathrm{E\mathring{E}}\epsilon W.$$

Hence $\mathrm{H\mathring{H}}$ is true $\iff \mathrm{H\mathring{H}}\epsilon W$.

COROLLARY 3.2. Letting F be the set of non-true sentences of $\boldsymbol{S}$, and $\overline{\mathrm{T}}$ the set of all expressions which are not true sentences, then neither $\eta(\overline{\mathrm{T}})$ nor $\eta(\mathrm{F})$ are definable in $\boldsymbol{S}$.

COROLLARY 3.3. If we extend $\boldsymbol{S}$ to a semantico-syntactical system $\boldsymbol{S}^C$, and if $\eta(\overline{\mathrm{Th}})$ are definable in $\boldsymbol{S}$, then $\boldsymbol{S}^C$ is semantically incomplete or inconsistent.

Actually, to apply Corollary 3.3 to concrete situations, one would most likely show that $\eta(\overline{\mathrm{Th}})$ is definable by showing (1) $\overline{\mathrm{Th}}$ is definable and (2) For any set W, if W is definable, so is $\eta(W)$. Semantical systems strong enough to enjoy property (2) (which is purely a property of $\boldsymbol{S}$, rather than of C) are of particular importance. We shall henceforth refer to such systems as semantically *normal* (or, more briefly, 'normal'). Thus $\boldsymbol{S}$ is normal if, whenever W is definable in $\boldsymbol{S}$, so is $\eta(W)$. Semantical normality is, of course, relative to the Gödel correspondence g.

COROLLARY 3.4. If $\boldsymbol{S}$ is semantically normal, then

(1) There is a Tarski sentence X for each definable set.
(2) F is not definable in $\boldsymbol{S}$, nor is $\overline{\mathrm{T}}$.
(3) If non-theoremhood of C is definable in $\boldsymbol{S}$, then $\boldsymbol{S}^C$ is semantically incomplete or inconsistent.

REMARK. The trivial systems $\boldsymbol{S}_P$ of Section 2 are semantically normal, relative to the correspondence g mapping each expression onto its quotation (the individual constants of $\boldsymbol{S}_P$ are, of course, the designators). In fact, Lemma 2.5 asserts precisely that. It is by virtue of normality that we showed the non-definability of $\overline{\mathrm{T}}$ in $\boldsymbol{S}_P$. $\boldsymbol{S}_P$ was deliberately constructed with the view of establishing normality as simply as possible.

We now turn to a non-trivial system $\boldsymbol{S}_A$, for which we easily establish semantical normality.

4. Systems of Arithmetic. The first arithmetical system $\boldsymbol{S}_A$ which we consider is much like arithmetic in the first order functional calculus. We have numerals (names of numbers), numerical variables, the logical connectives (all definable from the primitive '$\downarrow$' of joint denial), identity,

and the primitive arithmetical operations of · (multiplication) and ¬ (exponentiation). We depart from the lower functional calculus in that, given a (well formed) formula F and a variable, e.g., 'x', we form the (class) abstract ⌜x(F)⌝, read 'the set of x's such that F.' We use abstracts to form new formulas in two ways, viz., (1) For a numeral N, ⌜x(F)N⌝ (read 'N is a member of the set of x's such that F,' or 'the set of x's such that F contains N') and (2) ⌜x(F_1) = x(F_2)⌝ (read 'the set of x's such that F_1 is identical with the set of x's such that F_2.') By (2) we easily define universal quantification thus: ⌜(∀x)(F)⌝ $\underset{\text{df}}{=}$ ⌜x(F) = x(x = x)⌝.

A formal description of $\boldsymbol{S}_A$ now follows.

Signs of $\boldsymbol{S}_A$: x, ′, (,), ·, ¬, =, ↓, 1. We call these signs 'S_1', 'S_2', ..., 'S_9' respectively.

Rules of Formation, Designation and Truth

1. A numeral (string of '1's) of length n, designates the positive integer n.
2. 'x' alone, or followed by a string of accents, is a variable.
3. Every numeral and every variable is a term.
4. If t_1 and t_2 are terms, so are ⌜(t_1) · (t_2)⌝ and ⌜(t_1)¬(t_2)⌝. If t_1 and t_2 contain no variables and respectively designate n_1 and n_2, then the above new terms respectively designate $n_1 \times n_2$ and $n_1^{n_2}$.
5. If t_1 and t_2 are terms, then ⌜$t_1 = t_2$⌝ is a formula, called an atomic formula. All occurrences of variables are free. If no variables are present, then ⌜$t_1 = t_2$⌝ is a sentence, and is a true sentence if and only if t_1 and t_2 designate the same number n.
6. If F is a formula, α a variable, then ⌜α(F)⌝ is a (class) abstract. No occurrence of α in ⌜α(F)⌝ is free. If β is a variable distinct from α, then the free occurrences of β in ⌜α(F)⌝ are those in F. If F contains no free variable other than α, then ⌜α(F)⌝ is called a *predicate*, and the *abstraction* of F.
7. If H_1 and H_2 are abstracts, then ⌜$H_1 = H_2$⌝ is a formula. The free occurrences of any variable α in this formula are those of H_1 and those of H_2.
8. If α and β are variables, F_1 and F_2 formulae, and if ⌜$\alpha(F_1) = \beta(F_2)$⌝ contains no free variables, it is a sentence. It is true if and only if, for every numeral N, the result F_1(N) of replacing all free occurrences of α in F_1 by N, and the result F_2(N) of replacing all free occurrences of β in F_2 by N, are equivalent in $\boldsymbol{S}_A$ [i.e., are either both true in $\boldsymbol{S}_A$ or neither one true].
9. For any predicate ⌜α(F)⌝ and numeral N, the expression ⌜α(F)N⌝ is a sentence (as well as a formula) and is true if and only if the result F(N) of replacing all free occurrences of α in F by N, is true.
10. If F_1 and F_2 are formulae, so is ⌜(F_1)↓(F_2)⌝. The free occurrences of any variable α are those of F_1 and those of F_2. If F_1 and F_2 are sen-

tences, then $\ulcorner(F_1)\downarrow(F_2)\urcorner$ is a sentence and is true if and only if neither F_1 nor F_2 is true.[10]

Note: The notation 'F(N)' of (8) or (9) will also be used for an arbitrary term t, not necessarily a numeral — i.e., $\ulcorner F(t)\urcorner \underset{\text{df}}{=}$ the result of substituting freely the term t for the free variable of F.

Gödel Numbering. For any expression E, let $\sigma(E)$ be the string of Arabic numerals obtained by replacing S_1 by the Arabic numeral '1', S_2 by '2', ..., S_9 by '9'. This string $\sigma(E)$ designates (in decimal notation) a number, which we will call $\boldsymbol{g}_0(E)$. We shall take for our Gödel number of E (written '$\boldsymbol{g}(E)$' or '$\mathring{E}$') the number $\boldsymbol{g}_0(E)+1$.

It will facilitate our exposition if we identify the numbers with the numerals (strings of '1's) which designate them in $\boldsymbol{S}_A$. Then we define the norm of E to be E followed by its own Gödel number.

Arithmetization of the norm function. The following extremely simple definition accomplishes all the arithmetization of syntax which we need:

Def. 1. $n(x) \underset{\text{df}}{=} x \cdot 10^x$.

Explanation. If x is the g.n. (Gödel number) of E, then $n(x)$ is the g.n. of the norm of E. Thus, for example, 37 is the g.n. of S_3S_6. The norm of S_3S_6 is $S_3S_6\underbrace{S_9S_9\ldots\ldots S_9}_{37}$, and its g.n. is

$$\underbrace{3699\ldots.9}_{37}+1 = 3700\ldots0 = 37 \times 10^{37}.$$[11]

Semantical Normality. We say that an expression E satisfies the predicate H (relative to the Gödel correspondence g) if $H\mathring{E}$ is true. The set W of all expressions satisfying H is precisely the set defined by H, in the sense of Section 3. We also say that E satisfies the *formula* F (when F contains one free variable) if $F(\mathring{E})$ is true, and we shall also refer to the set W of all E which satisfy F as the set defined by the formula F. Now, the crucial role played by the class abstractors of $\boldsymbol{S}_A$ is that definability by a predicate, and definability by a formula, are thereby equivalent. This is an immediate consequence of Rule 9 of $\boldsymbol{S}_A$ since, if H is the abstraction of F, then E satisfies $H \Longleftrightarrow H\mathring{E}$ is true $\Longleftrightarrow F(\mathring{E})$ is true [by Rule 9] $\Longleftrightarrow$ E satisfies F. Thus the sets respectively defined by H and F are the same.

A formula F_N will be called a *normalizer* of formula F if F_N is satisfied by just those expressions E whose norm satisfies F. In the light of the preceding paragraph, the statement that $\boldsymbol{S}_A$ is semantically normal is

[10] We could have used the single primitive '$\subset$' [class inclusion] in place of the joint use of '$\downarrow$' and '$=$' (as occurring between abstracts). We would then have a system formulated in a logic based on inclusion and abstraction (in the sense of Quine). All results of this paper would still go through.

[11] Had we used $\boldsymbol{g}_0$, rather than $\boldsymbol{g}$ for our Gödel correspondence, then, if x were the g.n. of E, the g.n. of the norm of E would have been $(x+1)10^x - 1$, rather than $x.10^x$.

equivalent to the statement that every formula F (with one free variable) has a normalizer F_N (since F_N defines $\eta(W)$, when F defines W).

THEOREM 4. S_A is semantically normal, relative to g.

PROOF. We must show that every F has a normalizer F_N. Well, take F_N to be the result of replacing the free variable α of F by $\alpha \cdot 10^{\alpha}$ [or rather, by the unabbreviated form $\ulcorner(\alpha) \cdot ((1111111111\neg(\alpha))\urcorner$.]

Then, for any number x, $F_N(x)$ and $F(n(x))$ have the same truth-values. Thus, for any expression E,

$$\begin{aligned} E \text{ satisfies } F_N &\iff F_N(\mathring{E}) \text{ is true} \\ &\iff F(n(\mathring{E})) \text{ is true} \\ &\iff \text{the norm of E satisfies F (since } n(\mathring{E}) \text{ is the g.n. of the norm of E!).} \end{aligned}$$

Hence F_N is satisfied by those E whose norm satisfies F and is thus a normalizer of F.

COROLLARY 4.2. (1) For every definable set of S_A, there is a Tarski sentence. (2) The complement of the truth set T of S_A is not definable in S_A (relative to g). (3) T itself is not definable in S_A (relative to g). (4) Any proposed axiomatization of S_A such that the set of its theorems is definable in S_A (relative to g) is semantically incomplete or inconsistent.

(1) and (2) immediately follow from the preceding theorem, together with the results of Section 3. In particular, in (1), to construct a Tarski sentence for a set W defined by formula F, we first construct the normalizer F_N of F by the method of the preceding theorem, then take the abstraction H of F_N, and then follow H by its own Gödel number. Thus the Tarski sentence for W is the norm of the abstraction of the normalizer of the formula which defines W. (3) and (4) follow, since S_A contains negation (definable from '$\downarrow$').

REMARK. (4) of Corollary 4.2 can be thought of as one form of Gödel's theorem. Definability in S_A is actually equivalent to definability in protosyntax (in the sense of Quine). Thus any formal system for S_A whose set of theorems is protosyntactically definable will be semantically incomplete or inconsistent. This is essentially similar to Quine's result that protosyntax itself is not protosyntactically completable.

4.3. We have just shown a method for constructing normalizers which works for the particular Gödel correspondence g, which we employed. Actually, it will work for any Gödel correspondence relative to which the norm function (i.e., the function which assigns to each expression its norm) is *strictly* definable, in the following sense:

A function f (from expressions to expressions) will be said to be *strictly* defined by the term t (with one variable α), if, for any two expressions E_1 and E_2, $E_1 = f(E_2)$ if and only if the numeral $\mathring{E}_1$ and the term $t(\mathring{E}_2)$ [viz., the result of substituting $\mathring{E}_2$ for α in t] designate the same number.

This notion of strict definability is quite different from the usual much weaker notion of definability of f, viz., the existence of a formula M with two free variables such that, for any E_1 and E_2, $E_1 = f(E_2)$ if and only if $M(\mathring{E}_1, \mathring{E}_2)$ is true. We can, in an obvious manner, extend both notions of definability to functions of more than one argument.

If now the norm function is *strictly* definable relative to g, then to construct a normalizer for F we simply replace all free occurrences of the free variable α of F by the term $t(\alpha)$ which defines the norm function. Since this process nowhere makes use of quantifiers (or other identity of class abstracts),[12] or logical connectives, or more than one variable, then if we completely stripped $\boldsymbol{S}_A$ of its logical connectives, quantifiers, and all variables but one, the resulting vastly weaker system $\boldsymbol{S}_a$ would still be normal and, moreover, so would any extension of $\boldsymbol{S}_a$. Let us state this more precisely:

By the system $\boldsymbol{S}_a$, we mean the system whose signs are those of $\boldsymbol{S}_A$, except for '↓' and ''', and whose rules are those of $\boldsymbol{S}_A$, with the omission of Rules (7) and (10), and with Rule (2) changed to (2'), 'x' is a variable. By an extension of $\boldsymbol{S}_a$, we mean a system constructed from $\boldsymbol{S}_a$ by possibly adding additional signs and rules. Then the following theorem is a considerable strengthening of Theorem 4.1:

THEOREM 4.4. Any extension $\boldsymbol{S}'_a$ of $\boldsymbol{S}_a$ is normal relative to any Gödel correspondence g, relative to which the norm function is strictly definable providing that whenever two terms t_1 and t_2 have the same designata, $F(t_1)$ and $F(t_2)$ have the same truth-values.

4.5. NORMALITY OF $\boldsymbol{S}_A$ RELATIVE TO OTHER GÖDEL CORRESPONDENCES. If the norm function is definable (relative to g) in only the weaker sense, rather than strictly definable, then, although the above method of constructing normalizers is no longer available to us, we still have another method which will work for $\boldsymbol{S}_A$, but *not* for $\boldsymbol{S}_a$ (or any arbitrary extension thereof), since the construction depends on quantification.

Letting $N(\alpha, \beta)$ be the formula which defines the norm function, we let $F_N \overline{\overline{\text{df}}} \ulcorner(\exists\beta)(N(\beta, \alpha) \,\&\, F(\beta))\urcorner$, where the existential quantifier is defined from the universal quantifier in the usual manner, the latter defined as previously indicated, and '&' is defined from '↓' in the usual manner. Then F_N is a normalizer of F. Hence,

THEOREM 4.6. If the norm function is definable in $\boldsymbol{S}_A$ relative to a Gödel correspondence g, then $\boldsymbol{S}_A$ is normal relative to g.

We lastly observe that if there is a formula $C(\alpha, \beta, \gamma)$ such that, for any expressions E_1, E_2, E_3, $E_3 = E_1E_2$ if and only if $C(\mathring{E}_1, \mathring{E}_2, \mathring{E}_3)$ is true (which we express by saying that concatenation is definable, relative to g) and if

[12] As indicated at the beginning of Section 4, quantification is defined, using a formula which employs the identity sign between class abstracts.

there is a formula $G(\alpha, \beta)$ such that, for any expressions E_1 and E_2, E_1 is the Gödel numeral of E_2 if and only if $G(\mathring{E}_1, \mathring{E}_2)$ is true (which we express by saying that g itself is definable relative to g), then the formula $\ulcorner \exists(\gamma)(G(\gamma, \beta).\&.C(\beta, \gamma, \alpha))\urcorner$ defines the norm function and $\boldsymbol{S}_A$ is normal. Hence

THEOREM 4.7. A sufficient condition for $\boldsymbol{S}_A$ to be normal, relative to g, is that concatenation and g itself both be definable, relative to g.

REMARK. Gödel correspondences satisfying the hypothesis of Theorem 4.7 include all those that are *effective* (i.e., include all those g such that the function h, which assigns to each number x the Gödel number of (the numeral designating) x, is a recursive function.[13] This, in conjunction with previous results, yields the proposition that, relative to any effective Gödel correspondence g, the truth set of $\boldsymbol{S}_A$ is not definable. This, in essence, is Tarski's Theorem.

5. Concluding Remarks: Diagonalization vs. Normalization. We should like, in conclusion, to compare the norm function, used throughout this paper, with the more familiar diagonal function, used for systems in standard formalization.

Firstly, to sketch a general account of diagonalization,[5] analogous to Section 3 for normalization, we consider now an arbitrary language $\boldsymbol{L}$ which (like $\boldsymbol{S}$ of Section 2) contains expressions, sentences, true sentences, and individual constants. Instead of predicates, however, we now have certain expressions called 'formulas' and others called 'variables,' and certain occurrences of variables in formulas termed 'free occurrences,' subject to the condition that the substitution of individual constants for all free occurrences of variables in a formula always yields a sentence. We again have a Gödel correspondence mapping each expression E onto an individual constant $\mathring{E}$. For any formula F with one free variable α and any expression E, we define F(E) as the result of substituting $\mathring{E}$ for all free occurrences of α in F. The expression F(F) is defined to be the *diagonalization* of F. The set of all E such that F(E) is true, is called the set *defined* by F. For any set W, we define $D(W)$ as the set of all F whose diagonalization is in W. Then the analogue of Theorem 3.1 is 'A sufficient condition for the existence of a Tarski sentence for W is that $D(W)$ be definable.'

[13] E.g., the correspondence $\boldsymbol{g}_0$. To show that, relative to $\boldsymbol{g}_0$, the norm function is weakly definable, we must construct a formula $\ulcorner \varphi(\alpha, \beta)\urcorner$ such that, for any two numbers n and m, $\varphi(n, m)$ is true $\iff m+1=(n+1).10^n$ (cf. (9)). We first define addition as follows: Add $(\alpha, \beta, \gamma) \underset{\text{df}}{=} n^\alpha.n^\beta=n^\gamma$ (where we take n any number $\neq 1$). Then we define $\varphi(\alpha, \beta) \underset{\text{df}}{=} (\exists\gamma)$ [Add $(\alpha, 1, \gamma)$ & Add $(\beta, 1, \gamma.10^\alpha)$] (this construction can be simplified by introducing descriptors). Thus the tricky correspondence $\boldsymbol{g}_0+1$, which we used, was introduced only for purposes of simplicity, and is certainly not necessary for the success of our program.

Hence also, $D(\overline{W})$ is not definable. We would then define normality, for such a language $\boldsymbol{L}$, by the condition that whenever W is definable in $\boldsymbol{L}$, so is $D(W)$. Then all other theorems in Section 3 have their obvious analogues.[14]

To apply these general notions to systems in standard formalization, e.g., elementary arithmetic, we would have, in analogy with the notion 'normalizer,' that of 'diagonalizer,' where a diagonalizer F_D of a formula F would be a formula satisfied by just those expressions whose diagonalization satisfied F. Then, if W is defined by F, and if there exists a diagonalizer F_D for F, then the diagonalization of F_D (which is $F_D(F_D)$), is the Tarski sentence for W.

This is essentially the classical construction. The construction of the diagonalizer F_D is considerably more involved than the construction of the normalizer F_N. Again, we might say, this is due to the fact that concatenation is easier to arithmetize than substitution.

DARTMOUTH COLLEGE

[14] We can profitably avoid repetition of analogous arguments for $\boldsymbol{S}$ and $\boldsymbol{L}$ by regarding both as special cases of a more general structure. This approach will be presented in a forthcoming paper, *Abstract structure of unsaturated theories*, in which we study, in considerable generality, the deeper properties of undecidable systems, uncovered by Gödel and Rosser.

ON A FAMILY OF PARADOXES

A. N. PRIOR

1. Some paradoxical statements are, on the face of it, awkward for the propounder only, while some are also awkward for the looker-on. The Eubulidean version of the Liar paradox is of the second sort—if a man says 'What I am now saying is false', not only he himself but we who look on seem forced to say contradictory things (that his statement must be true because even if it were false it would be true, *and* that it must be false because even if it were true it would be false). On the other hand, if Epimenides the Cretan says that nothing said by a Cretan is the case, it appears that he has landed *himself* in a hole, but the beholder can contemplate his position without unease, simply saying that what Epimenides says must be false because even if it were true it would be false, and so concluding that it *is* false without further ado.

2. Church, however, has pointed out that there is a *little* further ado for the beholder nevertheless. For if what Epimenides says is false, then its contradictory, i.e. that *something* said by a Cretan is the case, must be true, and as the only Cretan statement we have been told about is false, this true Cretan statement which there must be, must be some other one than this. In other words, this one Cretan statement cannot even be made unless some other Cretan statement is made also.

3. Let us try formalising this proof in the propositional calculus enriched by (*a*) variables standing for monadic proposition-forming 'functors' of propositions (we shall use the one variable '*d*' for this purpose), and (*b*) quantifiers binding variables of any categories. We shall use *U* for the universal quantifier and *E* for the existential; for the rest Łukasiewicz's symbols, with *Q* for material equivalence as in *Aristotle's Syllogistic*. For postulates: substitution for variables (with the usual restrictions in the presence of quantifiers) detachment, Łukasiewicz's rules for the quantifiers, definitions of the various truth functions in terms of *C* and *U* (*Np* = *CpUpp*), and the one axiom *CCCpqrCCrpCsp*. This gives the full ordinary propositional calculus, but does *not* give any laws like *CdpCdNpdq*, *CQpqCdpdq*, *CddUppdp*, which in effect restrict the values of *d* to truth-functors. *d* can thus be used to stand for, among other things, the functor 'It is said by a Cretan that—', and where it occurs in the proofs below as a

Received August 3, 1960

free variable it will be helpful to assign this value to it illustratively. Note that allowing this as a value of d involves the view that 'It is said by a Cretan that p' is not a sentence about the sentence 'p' but a new sentence which, like 'Not p', is about whatever 'p' is about; e.g. 'It is said by a Cretan that Socrates is ill' is not about the sentence 'Socrates is ill' but is another sentence which, like that one, is about Socrates.

4. We can now give the following sketch proof of Church's conclusion as informally derived in 1 and 2:–

T1. $C(UpCdpNp)\,C(dUpCdpNp)\,(NUpCdpNp)$ – from $CUpdpdq$ by substitution.
T2. $C(dUpCdpNp)\,C(UpCdpNp)\,(NUpCdpNp)$ – from *T1* and $CCpCqrCqCpr$.
T3. $C(dUpCdpNp)\,(NUpCdpNp)$ – from *T2* and $CCpCqNqCpNq$.
T4. $C(dUpCdpNp)(EpKdpp)$ – from *T3* and equivalence of 'not-none' and 'some', i.e. of 'not-all-not' and 'some'.
T5. $C(dUpCdpNp)\,K(dUpCdpNp)\,(NUpCdpNp)$ – from *T3* and $CCpqCpKpq$.
T6. $CK(dUpCdpNp)(NUpCdpNp)(EpKdpNp)$ – substitution in $CdqEpdp$.
T7. $C(dUpCdpNp)(EpKdpNp)$ – syllogistically from *T5* and *T6*.
T8. $C(dUpCdpNp)\,K(EpKdpp)(EpKdpNp)$ – from *T4*, *T7* and $CCpqCCprCpKqr$.

5. What *T8* asserts, with our illustrative value for d, is that if it is said by a Cretan that whatever is said by a Cretan is not the case, then something said by a Cretan is the case, and something said by a Cretan is not the case. In order to pass from here to 'There are at least two statements said by a Cretan (or Cretans)' we need to introduce a functor Ipq for 'That p is the same thing as that q', either undefined with the two special axioms Ipp and $CIpqCdpdq$, or by definition as $UdCdpdq$, which will make these axioms theorems. We can then define 'dp for at least two p's' (put $E(2+)pdp$ for this) as short for $EpqKKdpdqNIpq$, and proceed thus: –

T9. $CIpqCKdppKdqq$ ($CIpqCdpdq$, subst.)
T10. $CIpqCKdppNKdqNq$ (*T9*, $CCpCqKrsCpCqNKrNs$).
T11. $CKKdppKdqNqNIpq$ (*T10*, $CCpCqNrCKqrNp$).
T12. $CEpqKKdppKdqNqEpqKKdppKdqNqNIpq$ (*T11*, $CCpqCpKpq$, quantification theory).
T13. $CEpqKKdppKdqNqE\,(2+)pdp$ (*T12*, *Df.* 2+).
T14. $CKEpKdppEqKdqNqEpqKKdppKdqNq$ (subst. in $CKEpdpEqgqEpqKdpgq$).
T15. $CKEpKdppEpKdqNqE(2+)pdp$ (*T14*, *T13*, *syll.*).
T16. $CdUpCdpNpE(2+)pdp$ (*T8*, *T15*, *syll.*).

And this is what we want–'If it is said by a Cretan that whatever is said by a Cretan is not the case, then at least two things are said-by-a-Cretan'.

6. If d is confined to truth-functors there is a shorter proof of *T8*, viz. this: There are only four monadic truth-functors, $V(Vp{=}Cpp)$, $S(Sp{=}p)$, N and $F(=NV)$. $SUpCSpNP$ $(=UpCpNp = UpNp)$, $NUpCNpNp$ and $FUpCFpNP$ are all clearly false, so for these substitutions the antecedent of *T8* is false and the whole true. $VUpCVpNP$ is true, but so are both $EpKVpp$ (provable from $KVVpVp$), and $EpKVpNp$ (provable from $KVFpNFp$) so for this value of d both antecedent and consequent of *T8* are true. For the rest, if d is confined to truth-functors, Ipq or $UdCdpdq$ is just Qpq and

$E(2+)pdp$ not only logically implied by but logically equivalent to $EpqKKdppKdqNq$, and if we used this form to define $E(2+)pdp$ the step from *T8* to *T16* would be immediate.

7. In fact, however, in proving *T8* and *T16*, I have *not* made use of any of those methods which would be available if *d* were confined to truth-functors. In this respect the proofs in 4 and 5 are like those used by Tarski to establish the equivalence of *Kpq* to the *first* and more complicated of his formulae with no constants but *Q* and *U*, and unlike the proofs he uses to establish the equivalence of *Kpq* to his second and simpler formula. (See his *Logic, Semantics and Metamathematics*, Paper I, and my critical notice of this work in *Mind*, July 1957, pp. 401-3.) In 4 and 5 nothing whatever is assumed about the possible values of *d* except that they must be functors constructing a statement out of a statement; if there are any other such functors beside truth-functors, *TT1-16* hold for them; and in particular if 'It is said by a Cretan that—' is such a functor, *TT1-16* hold for that functor too. At the same time, nothing is assumed in 4 and 5 that is peculiar to *non*-extensional functors or, e.g. to ones involving the notion of assertion; *TT1-16* apply to truth-functors too, also to modal functors (if these construct statements out of statements), and to ones involving not only the notion of assertion but also those of believing, hoping, fearing, etc. (under the same proviso). The following nice example of these other possible interpretations of *T16* is due to P. T. Geach: If it is feared by a schizophrenic that nothing feared by a schizophrenic is the case, then there must be at least one other schizophrenic fear beside this one. And the possibility of transposing our whole discussion into such terms as these has at least this importance: There is some temptation to argue that the functor 'It is said that—' takes as its argument not a sentence but the name of one, so that 'It is said that Peter is ill' is about the sentence 'Peter is ill' rather than about Peter; but there is surely not even a superficial plausibility in saying that 'It is *feared* that Peter is ill' is about the sentence 'Peter is ill' rather than about Peter, i.e. no plausibility in saying in this case that the subordinate sentence is being *mentioned* rather than being *used* (in the way that subordinate sentences are).

8. Geach has also pointed out that similar consequences follow from supposing that it is said by a Cretan (feared by a schizophrenic, possible, not the case, etc.), not that *nothing*, but just that *not everything* that is said by a Cretan (feared by a schizophrenic, etc.) is the case. This more modest assertion by a Cretan that *not everything* said by Cretans is the case, i.e. that at least *something* said by a Cretan is not the case—this is not, like the more sweeping Cretan assertion considered earlier, something that one cannot consistently suppose true. It is, however, something that one cannot consistently suppose false; for if it were false that some Cretan assertions are false, the truth would then be that no Cretan assertions are false, and so not this one either. But what this true assertion says is that at least one Cretan assertion is false; this cannot be the Cretan assertion we know about, for that one is *not* false, so if this Cretan assertion is so much as *made* (not only 'if it is true'—if it is made, it *is* true), there must be some other Cretan assertion beside it.

9. I want now to emphasise the *limited* character of what has been demonstrated so far. It has *not* been shown to be categorically impossible that a Cretan should ever say that nothing (or that not everything) said by a Cretan is true. What has been proved is *not* the categorical impossibility of anything in the nature of self-reference in assertions, fears, etc. All that has been proved is a *hypothetical* impossibility—what we have, if we apply the law of transposition to the theorems as stated above, is that *unless something else* is said by a Cretan (feared by a schizophrenic, possible, false, etc.) it cannot be said by a Cretan (feared by a schizophrenic, etc.) that nothing (or not everything) that is said by a Cretan (feared by a schizophrenic, etc.) is the case. If nothing else is said by any Cretan (even Epimenides) then indeed it is impossible for Epimenides the Cretan to say that nothing said by a Cretan is the case; whatever noises he makes, he will not under those circumstances be able to say *that* by them; though oddly enough the thing itself—that nothing said by a Cretan is the case—will under those circumstances be true, simply because there will under those circumstances be no Cretan assertions at all. If there *are* other assertions by Cretans, but all of them false, it will still be true, but still not sayable by a Cretan, that nothing said by a Cretan is true. That is, there will still be nothing wrong with *what we suppose the Cretan to say*, but only with *the supposition that* he says it. If on the other hand, there is at least one true assertion by a Cretan, it *will* be possible, at least as far as the above reasoning goes, for Epimenides to say that nothing said by a Cretan is the case, though of course this statement will then be a false one.

10. In fact we have proved in 4-5 nothing more than some simple corollaries of the obvious truth that *if it is a fact that no fact is asserted by a Cretan, then THIS fact (that no fact is asserted by a Cretan) is not asserted by a Cretan either.* Symbolically we might prove this thus:—

T17. $C(NEpKdpp)C(dNEpKdpp)(KdNEpKdppNEpKdpp) - CpCqKqp$, subst.
T18. $C(KdNEpKdppNEpKdpp)(EpKdpp) - CdqEpdp$, subst.
T19. $C(NEpKdpp)C(dNEpKdpp)(EpKdpp) - T17, T18, CCpCqrCCrsCpCqs.$
T20. $C(NEpKdpp)C(NEpKdpp)(NdNEpKdpp) - T19, CCpCqrCpCNrNq.$
T21. $C(NEpKdpp)(NdNEpKdpp) - T20, CCpCpqCpq.$

Similarly (corresponding to 8), if no falsehood is asserted by a Cretan, then the falsehood (as it will then be) that some falsehood is asserted by a Cretan, cannot be asserted by a Cretan either.

11. These limitations to what can be proved by the methods of 4 and 5 are, I think, to be welcomed rather than regretted. That nothing said by a Cretan is the case, is something that is in fact false; it is, however, logically conceivable that it should be true; and in either case, it is something that can be said. And to say that it could not under any circumstances, even the actual ones, be said by a Cretan, would surely be to put Cretans at an excessive disadvantage beside the rest of mankind. That our theorems stop short of this extreme seems therefore a recommendation of our logic. There are other points, however, at which the limitations

of our methods appear as odd and unpleasant gaps which cry out for filling up, if need be in some other way.

12. Let me turn at this point to a slightly more complicated case than any we have so far considered. L. J. Cohen, in the *Journal of Symbolic Logic* for September 1957, invites us to consider a policeman who testifies that nothing the prisoner says is true, while the prisoner says that something the policeman says is true. It is clear in the first place that the policeman cannot be right, for if (as the policeman avers) nothing the prisoner says is true, then the prisoner must speak falsely in saying that something the policeman says is true, and the truth must be that *nothing* the policeman says is true, and so not that thing either. But since what the policeman says—that nothing the prisoner says is true—thus implies its own falsehood, and is false, the truth must be that *something* the prisoner says is true. Now either this true thing the prisoner says is the statement we know about—that something the policeman says is true—or it is something else. If it is something else, the prisoner says something else. If not—if the prisoner's true statement is his statement that something the policeman says is true—then the *policeman* must say something else, for the only statement of the policeman that we know about *isn't* true. So we have this now proved: If the policeman and the prisoner make the two statements mentioned by Cohen, then at least one of them must say something else besides. Once again, there is no question here of a proof that the policeman and the prisoner categorically cannot make the pair of statements mentioned; all that is proved is that *if* neither of them says anything else, then necessarily *either* the policeman does not say that nothing that the prisoner says is true *or* the prisoner does not say that something that the policeman says is true. And I want to draw attention now not to the condition but to the disjunctive character of what comes after it. Our logic does not provide any means of deciding *which* of the two statements is precluded, or of proving that *both* are. And some may feel that this is an undesirable lacuna; and may feel this still more strongly after considering some allied cases.

13. In the Middle Ages the following puzzle was propounded by Jean Buridan (I modify slightly his example, which was passed on to me by P. T. Geach): Suppose there are four people who on a certain occasion say one thing each. A says that 1 and 1 are 2—a truth. B says that 2 and 2 are 4—a truth. C says that 2 and 2 are 5—a falsehood. And D says that exactly as many truths as falsehoods are uttered on this occasion. But if what D says is true, that makes 3 truths to 1 falsehood, so that it is false; while if it is false, that makes two truths and two falsehoods, and it is true. This reasoning can, I am sure, be formalised by the method of 4 and 5 into a proof of the following theorem: If not more than one thing is said by each of A, B, C and D on a certain occasion, and no one else says anything, then if A says that if p then p and B says that if p then p and C says that both p and not p, then D cannot say that exactly as many people speak truly as speak falsely on this occasion.

14. Note again that there is no question of proving that D categorically

cannot say the thing ascribed to him. The impossibility only arises *if* A, B and C say certain things. If, for example, all three of them say that *Cpp* (or *UpCpp*), there is no reason at all why D should not say that exactly as many people speak truly as falsely on this occasion, though under these circumstances such a statement would be clearly false. However, since what we have here is a thesis of the form *CpCqCrCsNt*, this says no more and no less than *CpCtCsCrNq*, i.e., if not more than one thing is said by A, B, C and D on a certain occasion, and no one else says anything, then if D says that exactly as many people speak truly as falsely on this occasion, C that (for some p) both p and not p, and B that (for all p) if p then p, then A cannot say that if p then p. In other words, D's saying what is attributed to him is not *more* blocked, as far as this logic goes, by the sayings of A, B and C than *their* sayings are blocked by what D is supposed to say; and if you hear all these four people together and then ask yourself 'Which of them is it who hasn't really said anything?', there is no more reason for answering 'D' than there is for answering 'A', 'B' or 'C'. For all that this fragment of logic has to tell us, it might just be a matter of who gets his say in first—if B, C and D really have said the things attributed to them, and it may be that they really can do this if they are quick enough, then A cannot say on that occasion that 2 and 2 are 4, or that if p then p. I must confess to a feeling that I would like to see a little more favouritism here; but I am not at all clear as to where it is going to come from, and I am not sure that the feeling isn't anyhow just a prejudice born of too much reading of *Principia Mathematica*.

15. It has been suggested (by B. Sobociński) that D's utterance must be separated from the rest as being in a 'different language' from them. On this view, D's 'truly' and 'falsely' are ambiguous expressions, and if L is the language of A, B and C, and D means 'truly-in-L' and 'falsely-in-L', then his own language cannot be L but must be some other. Hence he cannot himself be counted either among those who speak truly in L or among those who speak falsely in L, and his statement, though false in his own language, is not false-in-L and is not and cannot be among the statements which it is itself about. To this I would reply that in the story as given, nothing whatever is said about the *language in which* A, B, C and D say or do not say the things attributed to them; nor is D depicted in the story as making any reference either to his own language or to that of the others. And as for 'truly' and 'falsely', 'x says truly that p' is to be understood throughout as simply short for 'x says that p, and p'; 'x says falsely that p' as simply short for 'x says that p, and not p'; 'y says that x says truly that p' for 'y says that both x-says-that-p and p'; 'y says that x says something true' for 'y says that for some p, both x-says-that-p and p'; and analogously for 'y says that y says something true', and so on.

16. A language or languages *could*, however, be mentioned if one wished. In other words, the story of A, B, C and D *could* be re-told, and the associated theorem proved, with the simple 'says that' replaced by 'says-in-L that', where L is some specific language. That is, we can prove that if no more than these four things are said-in-L on this occasion,

then if A and B say-in-L that for all p if p then p, and C says-in-L that for some p both p and not p, then D cannot say-in-L that exactly as many people have spoken-in-L truly as have spoken-in-L falsely. And one could say, albeit a little loosely, that those who insist on a rigid hierarchy of 'language-levels' provide the 'favouritism' requested in 14, by asserting the final consequent of the preceding theorem categorically, i.e., in their metatheory of L they would say that whether A, B, and C speak as narrated *or not*, neither D nor anyone else can speak in L itself of what is said in L (whereas—this is where the favouritism comes in—they would *not* say that regardless of what D, C and B say-in-L, A cannot say-in-L that for all p, if p then p; at least, they would not say this if they supposed L rich enough to contain the propositional calculus with quantifiers). To this I would answer that (i) there are certainly languages L of which this last would be true, but (ii) it does not follow from this that there is no consistent language in which we can ever speak of what is said-in-any-language. This very thing, in fact, could not be said of the language in which it is said, if the advocates of rigid language-levels are right; and apart from that, there is no reason to suppose that the language used in this paper is inconsistent. What is true—and can be said in this language—is that there is no consistent language, and indeed no language at all, in which we can *always* speak of what is said-in-any-language.

17. Continuing the discussion of what cannot be proved by the methods of 3-5, it must be further mentioned that there is nothing in this part of logic to prevent Epimenides the Cretan from saying, regardless of what other Cretans say, that everything said by Cretans is *true*. And there is in fact no reason why he should not be supposed to say this, in the case in which there is some other Cretan statement which is false; for then this one would be obviously false also. But suppose there are no other Cretan statements but true ones. Would this one be true then? All but itself being favourably accounted for, whether *all* Cretan statements are true will depend on whether this one is. But whether this one is true depends on whether *all* are true, for that all are true is what it says. So we have an impasse—there is not and cannot be any reason for judging this assertion true rather than false, or false rather than true. And this seems to me sufficient reason for denying that this *could be said* by Epimenides under such circumstances. But there is no law in the system of 4 and 5 which would be instanced by 'If it is asserted by a Cretan that whatever is asserted by a Cretan is the case, then something asserted by a Cretan is not the case'. With the symbols available, the only such law could be $C(dUpCdpp)(EpKdpNp)$ – a principle which was suggested to me on these grounds by J. L. Mackie—but this is easily falsified by letting d be 'It is the case that—', making the whole equivalent to $CUpCppEpKpNp$, which has a logically true antecedent and a logically false consequent. (We could also falsify it by letting d be the modal functor L, 'Necessarily'.)

18. It seems similarly desirable to lay it down that if it is asserted by a Cretan that *something* asserted by a Cretan is the case, then something asserted by a Cretan must *be* the case. For if either nothing else were

asserted-by-a-Cretan at all, or there were other things asserted by Cretans but all of them false, then the truth or falsehood of this Cretan assertion that some Cretan assertion is true would depend on whether it was *itself* true (all other cases being non-existent or unfavourable), and this in turn (since *what is said* is that some Cretan assertion is true) would depend on whether *some* Cretan assertion is true, and we are infinitely see-sawing again. Yet we cannot derive the principle mentioned at the beginning of this section as an instantiation of $C(dEpKdpp)(EpKdpp)$; for with N for d this is a plain falsehood (likewise with the modal functor LN, 'It is impossible that–').

19. In the last two sections, although it is not possible to replace 'It is asserted by a Cretan that–' by certain truth-functors and modal functors, it *is* possible to replace it by 'It is feared by a schizophrenic that–' and other functors involving the notion of mental attitudes. And it may be that just as there are special laws (like the law of extensionality $CQpqCdpdq$) which fit truth-functors only, others which fit modal functors only, or only modal functors and truth-functors (all these being over and above what can be laid down or proved for all d-functors whatever), so there are special laws which only fit 'mental-attitude-functors', these special laws possibly including the pair mentioned in the last two sections.

20. It may also be observed that the counter-examples given to the formulae mentioned in 17 and 18 (call them 'Mackie's formulae') are ones in which the antecedent is a *necessary* truth (and the consequent *necessarily* false); and this may be of importance. Take the counter-example in 18, with $UpCpp$ for its antecedent. Why can we not proceed with $UpCpp$, 'Every proposition implies itself', as we did in 14 with the supposed Cretan assertion that every Cretan assertion is true? Why, that is, can we not say something like this: Every *other* proposition implies itself because in all the implications involved we have either antecedent and consequent both true or antecedent and consequent both false; so that leaves this proposition itself to consider; but how can we decide whether $UpCpp$ implies itself without *first* assigning *some* truth-value to $UpCpp$? Does not this land us in a circle as in the other case? No, because we know both that $UpCpp$ implies itself and that it is true because any proposition *must* imply itself–being a proposition necessitates self-implication. This solution is suggested by an early comment of McTaggart's on Wittgenstein (*Mind*, October 1923; *Philosophical Studies*, VIII); maybe it has a superstitious ring to contemporary ears, but I must say I would rather be suckled in this particular outworn creed than go back still further to the Ramified Theory of Types, which would at this point deny that $UpCpp$ itself was among the propositions substitutable for q in $CUpCppCqq$. Further, if one confined Mackie's formulae to cases in which the antecedent is contingent, this might turn out to tie up, in a contingent way, with the restriction suggested in 19; that is, it might turn out that the *only* contingent antecedents of the forms given ($dUpCdpp$ and $dEpKdpp$) are ones in which the d is a functor involving 'attitudes' like saying, thinking, hoping, etc.

21. What makes it a little odd that we cannot get what we want here by

the methods of 4 and 5, is that we *can* get things so very *like* what we want by those methods. For what our *T16* amounts to is that if a Cretan asserts that all Cretan assertions are false, what he says is not really falsifiable by purely 'internal' considerations alone. The self-refuting character of such an assertion can be offered as a *ratio cognoscendi* for its falsehood, but not as a *ratio essendi*. That its truth would entail its falsehood sufficiently proves that such an assertion must be false if it is made, but it cannot even be made unless there is some *other* reason for its falsehood than this one, namely a Cretan assertion distinct from itself which, being true, falsifies it by the straightforward method of being a counter-instance. Its self-refutation is only a *sign* that there must be a more straightforward refutation somewhere, if the thing is to be really asserted at all. And similarly with the self-confirmation of the weaker Cretan assertion that *some* Cretan assertion is false. It is interesting and perhaps even surprising that so much can be proved by so pure a logic as that used in 4-5; but what also seems surprising is that when this much can be thus proved, we cannot thus prove what is required for the cases considered in 17 and 18.

22. A further limitation to the logic of 4 and 5 may be noted in the following context: There can be a very great difference between the two forms *dEpdp* and *Epddp*. For it is quite certain that if anyone *says that there is something he is saying* (*dEpdp*) he cannot but be right; while it is strongly arguable that if there is anything *that a person says that he is saying* (*Epddp*), he cannot but be wrong. For the first: the theorem *CdEpdpEpdp*, 'If X says that there is something that X says, then there *is* something that X says' (namely *that*–that there is something that he says) is a simple substitution in *CdqEpdp*. For the second, we might begin from Geach's adaptation of a paradox of Buridan's: Suppose Simple Simon says 'I say that the earth is flat'; we reply 'It isn't'; and Simple Simon retorts 'I didn't say it was–I said *that I said* that it was'. If the 'I say' of S.S.'s first remark is performative rather than informative, his retort is false; but, Geach has pointed out (Buridan himself oddly failed to see this), if the retort is correct (as it is if the original 'I say' is informative) then the original statement is not, for in the original statement he says that he is saying that the earth is flat when in fact (as he himself points out in his retort) he is not saying that the earth is flat, but saying that he is saying that it is. This solution presupposes rather more than the apparatus of 3, 4 and 5, namely (a) that the proposition that someone says that *p* is always a different proposition from the proposition *p* itself (Simple Simon's saying that the earth is flat is a different thing–a different thing to assert, think, fear, etc.–from the earth's being flat) and (b) that anyone can only say one thing at a time. And by the methods of 3-5 we can prove the principle

$$C(UpNIdpp)\,C(UpqCKdpdqIpq)(UpCddpNdp),$$

which with the above (a) and (b)–which amount to the affirmation of the two hypotheses in the case in which *d* is 'X says at *t* that–'–will yield by detachment the conclusion that whatever anyone says at *t* that he says at *t*, he does *not* say at *t*, i.e. whatever anyone says at *t* that he says at *t*, he says falsely that he says at *t*.

23. But the main thing to be noticed with these theorems is not the need for the special hypotheses (a) and (b) in the case of the second one, but something which some people feel requires to be laid down even with the first one. Mackie has raised this point particularly in connection with the case of the theorem *CdEpdpEpdp* in which we let our *d* be 'Descartes thinks that—'. Descartes himself can be regarded as having argued in his *Cogito*, or in the patter which accompanies his *Cogito*, that if he thinks there is something that he thinks, then there *is* something that he thinks (namely, that there is something that he thinks), so that he cannot possibly be wrong about this. And I cannot see that this reasoning, so far as it goes, can be gainsaid. But Mackie suggests that a man cannot think *at all* (whether truly *or* falsely) that there is something that he is thinking, unless there is some other thing that he thinks besides. And I take it that anyone who agreed with this would also say that no one can say that he is saying that *p* unless he is also saying something besides this about what he is saying (a principle which with the postulate (b) of the last section would imply that no one can ever *say what he is saying* at all). I am myself inclined to think that the sort of self-confirmation and self-refutation involved in these cases is harmless. We are no doubt concerned here, as in 15 and 16, with talk about our talk (thinking about our thinking, etc.), but not, as in the earlier cases, with talk about the truth of our talk (though we draw conclusions about that). So it is not the truth of Descartes' thought that would give it its truth (as in the case in 15), or even its falsehood that would do so (as in the case in 8), but its very existence, i.e. its being thought; and 'self-confirmation' in this sense seems to me perfectly in order. But whether principles of the type suggested by Mackie are desirable or not, they are certainly not obtainable in the system 3-5.

24. As a formula embodying the principle to which Mackie is appealing here, Geach has suggested

(1) $C(Epdp)(EpKdpNIpEpdp)$,

i.e., *dp* holds for some *p*, only if there is a *p* other than the assertion itself that *dp* holds for some *p*, for which it holds; e.g., something is being thought, only if something other than that something is being thought, is being thought; and again, something true is being said by a Cretan, only if something true and other than that something true is being said by a Cretan, is being said by a Cretan. For its 'dual', Geach gives

(2) $C(UpCNIpUpdpdp)(Updp)$,

i.e., if *dp* holds for every *p* other than the assertion itself that it holds for every *p*, then it holds for every *p* absolutely. For example, if every Cretan assertion is true apart from a Cretan's assertion that every Cretan assertion is true, then every Cretan assertion, *simpliciter*, is true.

25. Are these formulae of Geach's open to objections similar to those which beset Mackie's formulae of 17 and 18? In the first place, we may note that the results of substituting *F*, *V* and *N* (or logically equivalent functors, i.e., *d*'s such that either *QdpFp*, *QdpVp* or *QdpNp* is a law of

the system) are provable even in our system of 3-5. (1) d/F and (2) d/V are settled by the mere fact that $EpFp$ is false and $UpVp$ true. Of the others, we may take as an example (1) d/N, i.e.

$$C(EpNp)(EpKNpNIpEpNp).$$

For this we simply prove in succession $EpNp$, $CIpEpNpp$, $CNpNIpEpNp$, $CNpKNpNIpEpNp$, and then our formula. The crucial case is therefore that in which we so substitute for d as to turn dp into the simple p (or into a formula logically equivalent to the simple p, e.g., NNp), i.e., we must consider particularly these two formulae:–

(3) $C(Epp)(EpKpNIpEpp)$

(4) $C(UpCNIpUppp)(Upp)$.

These are in fact independent of the basis in 3-5.

26. That they do not follow from this basis is clear from the fact that they are inconsistent with laws of extensionality, e.g., $CQpqIpq$, which are known to be consistent with that basis. For if we put Q for I in (3) and (4), as we would be entitled to do in a purely extensional system, we would obtain contradictions (since $QpEpp$ is logically equivalent to p and $QpUpp$ to Np). On the other hand, equally consistent with our basis in 3-5 (as may be shown by a simple four-valued matrix) are the two formulae

(5) $E(2+)pp$

and

(6) $E(2+)pNp$,

asserting that there are at least two distinct truths (i.e., p's such that p) and at least two distinct falsehoods (i.e., p's such that not-p), in the non-metalinguistic sense sketched in 5. And given these, it is not difficult to prove (3) and (4). For by (5), Epp is not the only true proposition, which (again interpreted non-metalinguistically) is essentially what the consequent of (3) asserts. And by (6), Upp is not the only false proposition, so that the antecedent of (4), which in effect denies this, is false.

27. There are thus no truth-functional counter-examples to Geach's formulae, as there were to Mackie's of 17 and 18. But might there not be others? Intuitively, the following case seems possible:– Let us suppose that all his life Mr. X was a great talker, and it became his ambition to talk his way right into the 21st century. Now picture him old and dying on the night of December 31st, in the year 1999. The clock is nearly pointing to midnight, and with his last breath the man says despairingly, "Everything said by me was-or-is said in the 20th century." But unknown to Mr. X, his clock was slow, and in fact the New Year had come in just before he spoke. It seems to me as obvious as anything of this sort can be, that this man's dying utterance was an honest error. But putting the functor "If it is said by Mr. X that–then it is said in the 20th century that–" for d in Geach's (2), it will assert that if everything *except* the assertion that everything said by Mr. X is said in the 20th century, is said in the 20th

century if said by Mr. X (and this is *ex hypothesi* the case), then *absolutely* everything said by Mr. X is said in the 20th century. From this it would seem to follow that what Mr. X said was actually *true*, which seems monstrous. It is not, indeed, quite as bad as that—what follows is rather that what would normally be expressed by Mr. X's last utterance is true, so that either he said something true by those words, or he said nothing at all by them, or something other than what would normally be expressed by them. This is bad enough; still, if the present paper shows anything it is that our intuitions in this area are not to be trusted, so Geach's formulae may well be laws nevertheless.

28. Summing up where we have got to so far: I have admitted that there are certain limits to the possibility of self-referring assertions, beliefs, fears, etc., some of these limits being established within a very general logic and some apparently requiring postulates of a more special sort. I have felt compelled, for example, to deny that a Cretan can assert either that all Cretan assertions are false or that all Cretan assertions are true (or that some are false or that some are true) if there are no other Cretan assertions of any kind. On the other hand I have insisted that even a Cretan *can* make these assertions if the conditions are favourable—if there is some other Cretan assertion and it is a true one, then even a Cretan may assert truly that some Cretan assertion is true or falsely that none is; and if there is some other Cretan assertion and it is a false one, then even a Cretan may assert truly that some Cretan assertion is false or falsely that none is. And there are some logicians who would say that here I am being less restrictive than I ought to be. What more do they want? and why?

29. The 'residual unease' which there may be at what I have said so far, has been expressed by J. L. Mackie in the following way: I have asserted in 17 that if a Cretan says that all Cretan statements are true, then we can look at other Cretan statements and if we find any of them false then the Cretan statement that all Cretan statements are true can be written down as *another* false Cretan statement, and that finishes the matter (though if all other Cretan statements were *true* we would perhaps be in a fix). But this, Mackie says, is as if we had a conjunction *Kpq* in which we first found *q* false, then on the strength of that found *Kpq* false, and could only give *p* any truth-value at all *after* we had taken these steps and assigned 'false' to the conjunction as a whole. And it seems incredible that the truth-value of a component of a conjunction should be thus determined by the truth-value of the conjunction as a whole. This argument, Mackie insists, does not depend on identifying the *meaning* of a universal proposition with the meaning of a conjunction of singulars (or of a singular and an exceptive). All he says is this: A Cretan's assertion that all Cretan assertions are true is true if and only if it is itself true and all other Cretan assertions are true. This looks as if its truth at least partly depends on its truth, and we know that this kind of 'dependence' does not admit of such reciprocation. When I argue that nevertheless its *un*truth need not thus depend on itself, for we could establish that solely on the grounds of the untruth of some *other* Cretan assertion, I am still making use of the above

if-and-only-if proposition, and moreover I am using it in a queer way—I get the falsehood of the original assertion from the falsehood of one member of the equivalent conjunction, and thereby and only thereby get the falsehood of the other member. For myself, I can only say that it seems that things *do* go this way sometimes.

30. Anyone sufficiently moved by the preceding argument may go beyond anything I would myself contend for, and hold that it is categorically impossible for a Cretan to make assertions about the truth-value of all or some Cretan assertions (in the sense which we who are not Cretans are able to give to 'all' and 'some'). But there are at least as good grounds for complaining that the system I have developed is *too* restrictive as there are for complaining that it is not restrictive enough; and it is to this new sort of complaint that I shall confine my attention from now on.

31. In 9, for example, and in 11-14, I have spoken freely about certain things being 'unsayable' in certain circumstances. Yet it seems quite obviously empirically possible that on a certain occasion (to take the example of 13) four persons A, B, C and D should respectively utter the sentence '*1* and *1* are *2*', '*2* and *2* are *4*', '*2* and *2* are *5*' and 'Exactly as many true things as false ones are being said on this occasion', and none of them utter anything further. (Cf. my review of Lewis Carroll, *Journal of Symbolic Logic*, September 1957, p. 310.) I am consequently committed to a distinction between the mere utterance of such sentences and actually saying what would normally be said by them. And independently of these puzzles there does seem to me to be everything to be said for making much a distinction. If Plato really says that Socrates is wise, then what we have here is not a relation between Plato and a sentence but one between Plato and Socrates; and whether Plato succeeds in thus relating himself to Socrates by relating himself in another way to a certain sentence, may well depend on all sorts of circumstances that we may take a while to notice. What turns out to be less straightforward than one might expect is the relation between the sort of thing done with functors like 'X says that—' in 3-5, and ordinary Semantics; I mean the kind of thing you get in Tarski's paper on Truth. In the system of 3-5, 'X says truly that *p*' can be defined very simply as 'X says that *p*, and *p*', so that we have it as a law that if X says that *p*, then he says so truly if and only if *p*. This, as far as it goes, is very like Tarski's 'Convention T' (*Logic, Semantics, Metamathematics*, pp. 187-8), though much simpler. But whereas the above rule is instantiated by

(a) *Whoever says that snow is white says so truly if and only if snow is white,*

the corresponding instantiation of Tarski's convention would be

(b) *The English sentence 'Snow is white' is true if and only if snow is white.*

Tarski's convention is not concerned with *saying truly that* something-or-other, but with the kind of truth that can be predicated of a form of words.

We might try to relate the two conceptions by equating (b) with

(c) *Whoever, speaking English, utters the sentence 'Snow is white', says something true thereby if and only if snow is white'.*

But this, I would contend, is false, and could only be derived from the true principle (a) by means of

(d) *Whoever, speaking English, utters the sentence 'Snow is white', thereby says that snow is white,*

which I should say is also false. For example, since snow *is* white, Tarski cannot say that snow is white by uttering either that sentence or any other, if he says nothing else immediately after I have said that either what I am then saying or what he will say immediately after, but not both of these things, is false (cf. 14). At least, (d) could fail for the assertion ascribed to myself. Though one does not need to drop (c) outright—one only needs to tack on to it 'provided that he does say *something* by it'. (d) also holds with the same proviso.

32. It is of some interest that, as Geach has pointed out, Jean Buridan was led by some of his paradoxes to a 'non-Tarskian' view of the language he was considering; in fact Buridan went much further in this direction than I would, abandoning even (a) above (or the principle that (a) illustrates). He argues, e.g. that there are circumstances in which the sentence 'What Plato says is false', uttered by Socrates, would be false even though what Plato says *is* false, and even though the same sentence, uttered by 'Robertus' on the same occasion, would not be false but true. He thinks this is what would happen if 'What Plato says is false' were the sole utterance of Socrates and 'What Socrates says is true' the sole utterance of Plato. I need not reproduce his argument; the case is clearly the same as Cohen's court-case, my own view of it being that under these circumstances at least one of the two philosophers would not succeed in saying anything at all, true *or* false, by his sentence. I would agree with Buridan that Robert could say something by uttering the same sentence on the same occasion, and this although Robert and Socrates utter it for the same reason, both falsely believing that what Plato is saying is 'God doesn't exist'. (This last subtlety is in the original.)

33. But do I in fact gain anything by the small pinch of non-Tarskianism that I have allowed myself? We must not forget how widely the variable *d* of our *TT 1-21* may range, and M. Dummett has pointed out that one of its possible values is 'Epimenides utters words which conventionally signify in his language that—'. Then we get, analogously to Geach's modification of the Epimenides in 8, the conclusion that unless someone (himself or another) has uttered *other* words which conventionally signify in his language something that isn't so, Epimenides cannot even *utter the words* which conventionally signify in his language that someone has uttered words which conventionally signify in his language something that isn't so. The answer to this, I suspect, is that *signifying that* something or other is not something that can be infallibly effected by our 'conventions'.

34. Even apart from this point of Dummett's, however, the distinction drawn in 25 between uttering such-and-such and *saying that* so-and-so, is only relevant in that very limited number of cases in which we let our *d* be 'It is said that–', 'It is denied that–', or some function of these. For example, what are we really supposing when we think we are supposing that some schizophrenic fears at *t* that nothing feared by any schizophrenic at *t* is the case, and that nothing else is feared by a schizophrenic at *t*? We are certainly not supposing that he utters words. Or again, take the following case: Mr. X, who thinks Mr. Y a complete idiot, walks along a corridor with Mr. Y just before 6 p.m. on a certain evening, and they separate into two adjacent rooms. Mr. X thinks that Mr. Y has gone into Room 7 and himself into Room 8, but owing to some piece of absent-mindedness Mr. Y has in fact entered Room 6 and Mr. X Room 7. Alone in Room 7 just before 6, Mr. X thinks of Mr. Y in Room 7 and of Mr. Y's idiocy, and at precisely 6 o'clock reflects that nothing that is thought by anyone in Room 7 at 6 o'clock is actually the case. Now in 4-5 it has been rigorously proved, using only the most general and certain principles of logic, that under the circumstances supposed Mr. X just *cannot* be thinking anything of the sort. What, then, *are* we in our muddle supposing him to be doing? Certainly something which to himself looking back on it a moment later would be quite indistinguishable from thinking that nobody in Room 7, etc. (and he might go home without ever learning of his error). How, we all want to cry out, can what a man is thinking and even what a man *can* be thinking on a given occasion, depend on what number is written on the other side of a door? That what a man can *truly* think should depend on things like this, is reasonable enough; but that what he can think *at all* should depend on such factors–can we swallow that? These cases are surely in a way *worse* than the simple Liar; for one might well agree that no man ever does just sit down and say (or think or fear) that whatever he says (or thinks or fears) is false; even the most stupid person must see that this is self-defeating and not do it without inserting or intending the obvious provisoes. But in the cases we are now considering, the things that we are supposing to be thought (feared, etc.) are things that quite easily could be thought (feared, etc.) by an intelligent and logically-instructed person, and that could even be thought (feared, etc.) by the very person we are puzzling about, if it were not for some quite contingent circumstance of which that person might well be for ever unaware.

35. It will not *quite* do to say that what is vexing us here is the idea that Mr. X could *think that he thinks* something when in fact he is not and cannot be thinking this thing. Indeed, if we suppose him to think that he thinks this thing at the same time as he thinks it, the situation (for the supposer) is vastly eased. For we can now suppose him to think falsely that no thoughts in Room 7 are true ones, this being false for the straightforward reason that *the thought that he is thinking that no thoughts in Room 7 are true ones* (which we now suppose him to be thinking as well as the other) is a true thought in Room 7. However, this is a somewhat special case, and if in our puzzle we replace thinking by fearing throughout, then

part of the puzzle would be that Mr. X (even supposing him the best introspector in the world) could think that he is afraid that nothing feared in Room 7 is the case, when in fact he is not and logically cannot be afraid of anything of the sort, and this not because of a logical but because of an empirical fact of which he happens to be ignorant. But with thinking, as we have stated the case in the last section, the difficulty is simply that to *us* something should appear to be quite obviously empirically possible which is in fact not even logically so.

36. It is rather tempting to say about the man in Room 7 that what we have misdescribed as a thought of his about Room 7 is in fact a thought of his about Room 6. But we would not consider ourselves justified in saying this in other cases which closely resemble the present one but do not happen to issue in paradoxes. Let us suppose, for example, that Mr. X has gone not into Room 7 but into Room 9, and knows perfectly well that that is where he is, but still thinks mistakenly that Mr. Y is in Room 7 when in fact he is in Room 6. We may again suppose him to think at 6 o'clock that nothing thought at 6 o'clock in Room 7 is the case; but now there is nothing at all contradictory in this supposing; we may even suppose him to think *rightly* that nothing thought in Room 7 is the case, this being true because although Mr. Y is not in Room 7 someone equally idiotic is (or perhaps because no one is). For such a case we would surely say that Mr. X was right about Room 7, though for wrong reasons. And if in fact the occupant of Room 7 was a perfectly sensible person whose thoughts at 6 o'clock were true ones, we would say that Mr. X had thought something about Room 7 that was wrong, rather than that he had not been thinking about Room 7 at all but about Room 6 (and so was actually *right* in what he thought). So I don't think this way out will do.

37. J. L. Mackie suggests that while we cannot deny the empirical possibility of Mr. X's thinking that nothing thought at 6 in Room 7 is the case, even under the circumstances envisaged in 34, the reasoning in that section shows that he cannot think this *non-paradoxically*, paradoxicality and its absence being features of thinking which are not always introspectible. But either Mackie's phrase 'paradoxical thinking' refers to some species of thinking or it does not (it would not if 'paradoxical' were an *alienans* adjective like '*soi-disant*'). If it does not, i.e. if paradoxical thinking is no more a kind of thinking than imaginary money is a kind of money, then the conclusion of the argument of 34 is admitted. If, on the other hand, paradoxical thinking *is* thinking, then that argument shows that under the circumstances described it cannot occur, i.e. Mr. X cannot think either paradoxically or non-paradoxically, in Room 7 at 6, that nothing thought in Room 7 at 6 is the case, if this is all that is thought (paradoxically or non-paradoxically) in Room 7 at 6; for if he did, it both would and would not be the case that nothing thought in Room 7 at 6 was the case. The trouble here is that if we suppose Mr. X to have this thought it is not merely Mr. X but *we* who 'think paradoxically', in the only too straightforward sense of contradicting ourselves; and the job of being rigorously rational even about

irrationality (which is surely what all this consideration of paradoxes is in aid of) is just not done.

38. Further, even if we take the line that Mackie's 'paradoxical thinking' is not thinking, while this provides at least a verbal solution to the case of Mr. X (more than verbal if we can see what, positively, this paradoxical-thinking *is*), it gives rise to analogous problems of its own; at least it does so if it makes sense to say that someone paradoxically thinks that *p*. For this then becomes a possible value of *dp* in 3-5, and we can show that no one can paradoxically-think that nothing that he paradoxically-thinks is the case, unless there is something else, and something that *is* the case, that he paradoxically-thinks as well.

39. At this point I must confess that all I can say to allay the misgivings expressed in the past four sections is that so far as I have been able to find out, my terms are the best at present offering. I have been driven to my conclusion very unwillingly, and have as it were wrested from Logic the very most that I can for myself and others who feel as I do. So far as I can see, we must just accept the fact that thinking, fearing, etc., because they are attitudes in which we put ourselves in relation to the real world, must from time to time be oddly blocked by factors in that world, and we must just let Logic teach us where these blockages will be encountered.

40. Look back again at the grand simplicities of 10, and apply them here. If it is a fact that *no* fact is being assented to in Room 7 at 6, then *this* fact (that no fact is being assented to, etc.) cannot be being assented to in Room 7 at 6. There just isn't any way round this, is there? Not, anyhow, unless one says with the Ramifiers that there is no such thing as a plain fact, but only first-order facts, second-order facts, and so on; that the fact that no first-order fact is being assented to in Room 7 is itself not a first-order but a second-order fact; and that the fact that no fact of *any* order is being assented to in Room 7 is not and cannot be assented to by anyone at all, even in Room 9, because there is not and cannot be any such fact. This would be to dispose of an argument for certain restrictions on what is allowed to be sayable, thinkable, etc., by admitting both these and countless other restrictions by another door; not, it seems to me, the shrewdest of bargains. One can admit, however, that it is when he is 'order-jumping' (or at least when someone in his neighbourhood is doing so) that the world's best introspector is liable to find himself deceived.

Manchester University
Manchester, England

Notre Dame Journal of Formal Logic
Volume V, Number 3, July 1964

A NOTE ON SELF-REFERENTIAL STATEMENTS

NICHOLAS RESCHER

The standard resolution of the semantical paradoxes arising from self-referential statements is to dismiss these statements *en bloc* as meaningless. In a recent article, A. N. Prior deplores this wholesale solution as too drastic, and urges a more selective procedure. ("On a Family of Paradoxes," *Notre Dame Journal of Formal Logic*, vol. 2 [1961], pp. 16-32.)

Prior's approach—if I understand him aright—is to dismiss as *prima facie* meaningless only those self-referential statements which cannot consistently be classified as either true or false. This includes not only the various well known semantical paradoxes such as that of the Liar, but also the following interesting case (due, in its essentials, to John Buridan of Buridan's Ass fame): Messrs. A, B, C, and D make statements on a certain occasion, A and B both uttering some palpable truth (say: $1 + 1 = 2$), C a palpable falsehood (say: $1 + 1 = 1$), and D saying that just as many speakers speak truly as falsely on this occasion. (Thus if D's statement is *classed* as true, he speaks a falsehood; and if it is classed as false, he speaks a truth). In such cases, Prior would reject the pivotal statements as meaningless specifically because they cannot viably be classed as true or as false—and *not* generically because they involve self-reference. Had Mr. "Liar" said that his (self-same) statement was *true*, then—since no impossibility inheres in classing *this* statement of his as true—Prior would (I take it) be prepared to accept the self-referential statement as meaningful. Or again, had Buridan's last speaker said that *fewer* truths than falsehoods were spoken on the occasion in question, his self-referential could be classed as false without giving rise to difficulties, and would thus be meaningful on Prior's criterion.[1] Prior's solution thus has the important merit

1. An interesting if not strictly relevant case arises if D says that a least three truths are spoken. For this statement—which could feasibly be classed as false—is self-validating: if taken as true it *is* true.

Received October 15, 1962

of liberality—it exiles self-referential statements from the domain of the meaningful not as a matter of inflexible policy, but only in cases of actual necessity.

One immediate—and of itself by no means unacceptable—consequence of this criterion is that not merely will certain self-referential statements be *meaningful,* but some of them will even have to be regarded as *necessary.* For example the statement "There are false statements," symbolically "$(\exists p)\sim p$," will have to be regarded as a necessary truth. (It cannot be classed as false, since it can be inferred from its own denial; on the other hand no difficulty ensues if it is accepted as true.)

So far so good. But now, as Prior points out, a further much more subtle complication must be introduced, namely that self-referential can be such that if certain preconditions fail to be satisfied these statements "cannot even be made" (in Prior's language) or rather, they are only *conditionally meaningful* (as I would prefer to put it).

Suppose that Epimenides the Cretan says that nothing said by a Cretan is the case. Then we could readily class Epimenides' statement as false, though it could not possibly be true. But this, as Church has pointed out, commits us to accepting as true the *contradictory* of this statement, viz. that something said by a Cretan is true. Now since the only Cretan statement we have been told about is false, this true Cretan statement—which we are thus committed to suppose—must be some other statement. Thus, if we are to regard Epimenides statement as meaningful (and thus false), we are committed to presuppose the existence of at least one true Cretan utterance [a contingent fact]. Epimenides' statement is thus—on the approach—only conditionally meaningful. It is indeed conditionally L-false. It will have to be classed as false whenever meaningfulness-condition is assumed to be satisfied.

Similarly—and somewhat more unpleasantly—it is easy to devise an example of a conditionally L-true statement. Suppose that Mr. X makes (in Noplacese) the statement that someone has (at some time or other) made a false statement (in Noplacese). We cannot possibly class this statement of X's as false, for in doing so we *eo ipso* render it true. Thus if we are to regard the statement as meaningful we must class it as true. But it then entails the existence of a Noplacese utterance distinct from itself (viz. one that is false). Therefore, if we are to regard Mr. X's statement as meaningful (and thus true) we are committed to presuppose the existence of at least one false Noplacese utterance (a contingent fact). Mr. X's statement is thus only conditionally meaningful, and is indeed conditionally L-true. It will have to be classed as true whenever its meaningfulness condition is assumed to be satisfied.

The disadvantage of Prior's approach is illustrated by these examples. In certain cases it leads to the consequence that there are statements whose very *meaningfulness* (and not merely truth or falsity) can hinge upon a matter of contingent fact. And moreover this contingent meaningfulness gives rise to the anomaly that there are *conditionally* L-true (and L-false)

statements—statements which in the very logic of things *could* not possibly be meaningless if some *purely contingent* precondition failed to be satisfied.

I confess to being much in sympathy with the spirit of Prior's approach of avoiding the somewhat Procrustean policy of dismissing self-referential statements *en bloc* as meaningless. Very possibly the advantages of such greater liberality could outweigh its having certain somewhat distasteful consequences must, however, be recognized.

University of Pittsburgh
Pittsburgh, Pennsylvania

Corrigenda:

Page 218, addendum to footnote 1: Again, note that with the pair (1) "(2) is false," (2) "(1) is false" we could indifferently class either as true and its partner as false.

Page 219, paragraph beginning "Suppose that Epimenides...," last sentence: Insert "its" between "whenever" and "meaningless-condition."

TOWARD A SOLUTION TO THE LIAR PARADOX[1]

by Robert L. Martin

I. Preliminaries

1. This very sentence is in English.
2. Sentence 2 is interesting.
3. This very sentence is about itself.
4. It is only a token of this sentence type that confronts you.
5. Everything I write, including this, is considered carefully.
6. This very sentence does not exist.

IN WHAT follows I shall propose a solution to the Liar Paradox that does nothing to impugn sentences of the sort listed above. On the contrary, it provides some explication of the legitimacy of "good" self-reference, by contrasting it with "bad" self-reference. What it means to say that some self-reference is bad is clear; contradiction follows.[2] What it means to say that some self-reference is legitimate is less clear: roughly, that sentences such as those listed above can be readily understood by a speaker of the language, and can be judged, in a logically unproblematic way, to be straightforwardly true (or false).[3]

[1] I gratefully acknowledge the help of M. Fisk and R. H. Thomason.

[2] This is *not* to say, at the outset, that " This very sentence is true, " from which no contradiction follows, is not bad. I take it that there are some clearly legitimate cases of self-referential sentences, and clearly illegitimate cases (the ones that lead to contradiction), which provide one adequacy criterion for a solution—that the solution follow from a theory which rules out the clearly bad cases and allows the clearly good cases. If this much can be achieved (subject, of course, to other adequacy criteria), then, I imagine, one should let the theory decide about the questionable cases. (Qualification: talk of adequacy criteria and theories suggests a more rigorous presentation than I am able to offer.)

[3] There is no novelty in saying that some self-reference is legitimate (Wittgenstein, *Tractatus* 3.332 notwithstanding). See, e.g., R. Carnap, *Logical Syntax of Language* (London, 1959), pp. 53 ff., and W. and M. Kneale, *The Development of Logic* (Oxford, 1962), p. 718. R. M. Martin has written: " so far as is known, a syntactically self-referent language L is harmless . . . self-reference of a semantical kind is in some sense too much and leads to contra-

It follows from the familiar levels-of-language approach to the paradox (often credited to Tarski, despite his explicitly negative conclusion about the possibility of avoiding contradiction in natural languages)[4] that *all* self-referential sentences are illegitimate (ill-formed). For to avoid being *ad hoc* such an approach is based generally on the distinction between the use and (any sort of) mention of expressions, and in all the sentences above (in all self-referential sentences) use and mention are fused as much as in the paradoxical "This very sentence is false."

"But something must be given up," it might be argued in defense of the language-level approach. "The language-level treatment of the Liar has the great advantages of generality and intuitive clarity. For it deals with *all* the semantic paradoxes, and is based on an intuitively clear distinction. The price we pay—regarding all self-referential sentences as ill-formed—is no more than one should expect. Remember, the semantic paradoxes necessitate *some* intuitively unanticipated restriction, just as Russell's paradox and others have necessitated restrictions on the notion of class existence."[5]

In answer to this, surely it is *not* clear that the price need be as high as the restriction against all sentences that mention themselves. For the solution offered here (which seems capable of extension to the other semantic paradoxes) we shall pay considerably less.

Notice that if the sentences displayed above do seem worth saving, as I shall henceforth assume, then one temptingly simple way of separating good from bad self-reference appears less than

---

diction. Some kinds of self-reference are therefore legitimate, whereas other kinds are intolerable. Precisely how one draws the dividing line between tolerable and intolerable kinds of self-reference is by no means clear at the present time" (*Truth and Denotation* [Chicago, 1958], p. 138).

[4] See Alfred Tarski, "The Concept of Truth in Formalized Languages," in *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, trans. by J. H. Woodger (Oxford, 1956), pp. 164, 165. W. Kneale, for example, seems to overlook this; he speaks of the language-level solution to the Liar as flowing directly from Tarski's own conclusions (*op. cit.*, p. 665). I do not wish to deny that the language-level approach to the semantical paradoxes is to a large extent based on Tarski's work.

[5] W. Quine makes this point in *Set Theory and Its Logic* (Cambridge, Mass., 1963), p. 255, but not in support of any particular approach.

satisfactory. It has been suggested that a sentence may talk about its own structure or shape, but not about its own meaning.[6] But sentences such as 2, 3, and 5 above certainly *may* be interpreted as being about their own meanings.

Before proceeding, we must ask: is there in fact *one* Liar Paradox? Many puzzles, differing in apparently nontrivial ways, have been called by this name,[7] and there is the possibility that a solution to one may not be a solution to another. I propose to deal with a highly simplified version. We can be sure, at least, that no solution which fails to treat the version I shall consider can be of much value. As to whether the solution I shall offer can be extended to handle the other versions satisfactorily—I shall try to show that this can be done. It may be well, even now, to discuss briefly the connection between what I will take to be the "crux" version and other more complicated versions.

I take it that all versions of the Liar involve some kind of self-reference, and something amounting to the attribution of falsity. This is most direct in the crux version:

This very sentence is false.

(where it is understood that the expression "This very sentence" refers to the whole expression of which it is a part.[8] But it seems clear that the effect is the same (a sentence manages to assert that it is false) in cases involving what I shall call "indirect self-reference":

The next sentence is true.
The preceding sentence is false.[9]

[6] See, e.g., J. L. Austin, "Truth," in his *Philosophical Papers* (Oxford, 1961), p. 94, n. 2, and, for a more explicit statement, Kneale, *op. cit.*, p. 719.

[7] There is an inventory of versions in Rüstow, *Der Lügner: Theorie, Geschichte, und Auflösung* (Leipzig, 1910), pp. 40 ff. See also J. Agassi, "Variations on the Liar's Paradox," *Studia Logica*, 15, (1964-1965), pp. 237-238.

[8] In some, perhaps most, contexts, even "This very sentence is false" would be taken to refer to some other sentence. It seems possible, however, to construct a sentence *all* of whose tokens would be self-referential: perhaps "This sentence, the one you are now in the midst of reading (or hearing), is false." Assuming that one could form such a sentence, we may, for convenience, continue to deal with "This very sentence is false" as if it were such.

[9] Of course neither sentence is, by itself, self-referential. (Carnap has a similar example to show that the paradox does not depend solely upon *direct* self-reference—see his *Logical Syntax*, p. 215.) I have found it useful to assume

as well as in cases involving what I shall call "empirical self-reference":

(1) Sentence 1 is false.

Again, there is self-referential attribution of falsity *involved* in some general sentences,[10] such as

Every sentence is false.

or

Every Cretan sentence is false [said by a Cretan].

The striking difference between the crux version and versions involving indirect self-reference, empirical self-reference, and generality of reference is this: versions of the latter sort involve sentences, only tokens of which are self-referential. I take it that every token of "This very sentence is false" is self-referential; on the other hand, it is clear that in some contexts, "The next sentence is false" does not refer to itself even indirectly. Similarly, it is only when a token of "Sentence 1 is false" happens to be numbered one that it refers to itself. In versions involving generality of reference, only some cases involve sentences only some of whose tokens are self-referential. For while "All Cretan sentences are false" (said by me) is not self-referential, any token of "All sentences are false" *does* involve self-reference.

It is because of this difference between kinds of versions of the Liar that we will distinguish between "semantical incorrectness"

---

that there is a kind of self-reference which is indirect (that is, a sentence may refer to itself indirectly, via other sentences, as well as directly). What is needed, of course, beyond this assumption, is an analysis of indirect self-reference which explains why reference here, but not in general, seems to be a transitive relation.

[10] The fifteenth-century Paul of Venice concurs: " of propositions having reflexion on themselves, some have the property that their significations terminate solely at themselves, e.g., ' This is true, ' ' This is false, ' indicating themselves. But others have the property that their significations terminate both at themselves and at other things, e.g., ' every proposition is true, ' ' every proposition is false. ' " See Bochenski, *A History of Formal Logic*, trans. and ed. by I. Thomas (Notre Dame, Indiana, 1961), p. 248. Paul of Venice also speaks of something close to what I have called indirect self-reference (" of propositions having reflexion on themselves some have this reflexion immediately, others have it mediately") and offers the following general definition of *insolubile* (the family name for versions of the Liar): " *propositio habens super se reflexionem suae falsitatis aut se non esse veram totaliter vel partialiter illativa.* " See Kneale, *op. cit.*, p. 228.

(applicable to sentence types) and "token oddity" (applicable to sentence tokens). The notion of token oddity, which will be used to solve versions involving indirect self-reference, empirical self-reference, and (some of those involving) generality of reference, will be explained in terms of semantical incorrectness, which is used to solve the crux version and some versions involving generality of reference. Let us proceed to an explicit formulation of the crux version of the paradox.

Consider the following expression, whose type we may conveniently denote with the sign "(a)":

This very sentence is false.

1. (a) is a sentence.
2. Every sentence is true or false, but no sentence is both true and false.
3. (a) is true or (a) is false, but not both.
4. Suppose (a) is true. If a sentence is true then what it says is the case. But (a) says that (a) is false. Therefore, it is the case that (a) is false.
5. Therefore, (a) is false.
6. (a) says that (a) is false. This is the case (5). If a sentence says what is the case, then it is true. Therefore, (a) is true.
7. (a) is both true and false (5 and 6) and not both true and false (3).

To avoid the contradiction, at least one of the steps 1 through 6 must be denied. I shall deny 2 on the grounds that not all sentences (grammatical strings) but only those that are semantically correct (in the sense to be discussed) are either true or false. It will be seen that (a) can never pass what I shall argue is a reasonable test of semantical correctness (can never be shown to be semantically correct), so that there is no reason to allow step 3.

The problem of solving the Liar is quite independent of the question of the proper bearer of truth and falsity. I shall be speaking of sentences being true or false, but this is quite inessential; I could as well speak of the propositions, or judgments, expressed by sentences as being true or false, or of the assertions or statements made with the help of sentences, as being true or false. (a) could be recast as the sentence "The proposition

expressed by this very sentence is false," step 1 as "(a) expresses a proposition," step 2 as "Every proposition is true or false but not both true and false," and so forth. Then the contradictory conclusion would be "The proposition expressed by (a) is both true and false and not both true and false." The line I should then take would be: deny step 1—that (a) expresses a proposition—by showing that only semantically correct sentences express propositions and that (a) cannot pass the test of semantical correctness. Similar recastings could be made to conform to the view that statements or assertions are what are true or false. It is surely more convenient to discuss the Liar in terms of the truth and falsity of sentences. Whoever urges that statements or propositions or judgments are *really* the bearers of truth and falsity should, after all, supply a precise account of the relation that holds between such things and the more everyday units of written and spoken language. We shall be spared that task.

Another word on neutrality. It has been argued by Black and Sellars that Tarski's semantical schema

(T) X is true if and only if p

(where the letter X is to be replaced by the name of a sentence, and the letter p by that sentence itself—or a translation of it into the metalanguage), has practically nothing to do with the philosophical dispute (on the nature of truth) among adherents of the correspondence, coherentist, or pragmatic theories.[11] If they are right, as it would seem they are, then the problem of the Liar is also independent of that philosophical controversy, for the schema (T) is all one *needs* (as far as the notion of truth is concerned) to formulate the Liar.[12] It is interesting to notice that

[11] See M. Black, "The Semantic Definition of Truth," *Analysis*, 8 (1948), reprinted in *Philosophy and Analysis*, ed. by M. Macdonald (Oxford, 1954), pp. 245-259, and W. Sellars, "Truth and 'Correspondence,'" *Journal of Philosophy*, 59 (1962), reprinted in Sellars, *Science, Perception and Reality* (New York, 1963), pp. 197-224 (see esp. pp. 197, 198). Tarski himself seemed to feel that his schema captured the essential tenet of the correspondence view of truth (see his "The Concept of Truth in Formalized Languages," pp. 153, 155).

[12] See, e.g., Tarski's formulation in "The Concept of Truth in Formalized Languages," *loc. cit.*, p. 158.

formulations of the Liar for which that schema is sufficient are what I have called empirical versions; the crux version of the Liar, formulated above, relies on a stronger schema:

(T′) X is true if and only if what X says is the case.

(where X is replaced by the name of a sentence) whose plausibility *does* seem to rest on the acceptance of something in the tradition of Aristotle's correspondence "formula" (which is stronger than the semantical schema):

> For it is false to say of that which is that it is not or of that which is not that it is, and it is true to say of that which is that it is or of that which is not that it is not [*Meta.*, 1011b 26-27].

A coherentist or pragmatist might deny that the truth of a sentence depends fundamentally on whether what it says is the case, but (so goes the argument) neither would take exception to the claim that, for example, "Snow is white" is true if and only if snow is white. Despite the fact that the version I will be dealing with first involves the stronger schema, I can justly claim neutrality on the philosophical dispute about truth. For the paradox seems to arise in any of the views, in some form with which I have promised to deal.

II. Grammatical and Semantical Categories: Semantical Incorrectness

The proposed solution is based on the doctrine of linguistic categories,[13] and I distinguish, to begin with, between grammatical and semantical categories. I am not sure that this is a very fundamental distinction (more on this below) but the distinction is fairly well established and useful for our purposes. Grammatical categories are called for in the explication of *grammatical expression* (sentence). What is to be accounted for is the difference between expressions such as "The book is on the table" and (grammatically deviant) expressions such as "Book the table the on is."

[13] For background notes on the doctrine, see Bochenski, "On The Syntactical Categories," *New Scholasticism*, 23 (1949), pp. 257-280. Reprinted in *Logico-Philosophical Studies*, ed. by A. Menne (Dordrecht, 1962), pp. 67-87.

The set of sentences of a language does not coincide with the set of nondeviant sentences of that language; it is this fact which suggests the need for a class of semantical categories, distinct from grammatical categories. To use Chomsky's example,[14] both

1. Colorless green ideas sleep furiously.
2. Furiously sleep green ideas colorless.

are deviant but only 2 is ungrammatical. Imagine that one had an ideal grammar. Would one expect this *grammar* to rule out "Golf plays John" while allowing "John plays golf"; to rule out "Sincerity admires Jane" while allowing "Jane admires sincerity"? Perhaps such distinctions could be made satisfactorily only on a semantical basis, taking into account not only the grammatical roles but also the *meanings* of the terms in question. "Sincerity admires Jane," one might say, is in accord with the rules of the grammar, but since sincerity is not the *kind of thing* that could admire, or fail to admire, anyone or anything, the expressions "sincerity" and "admires Jane" cannot be juxtaposed (that is, they belong to incompatible semantic categories). The category correctness defined by a grammar, one might conclude, is a necessary but not sufficient condition for a string of words being nondeviant; besides being grammatical, a nondeviant sentence is semantical-category correct.[15]

There are those, on the other hand, who feel that our ideal grammar could carry us further. Chomsky has written persuasively in favor of the notion of a "degree of grammaticalness" which would give a clear sense to the idea that "John plays golf" is *more* grammatical than, say, "Golf plays John," while the latter is in turn *more* grammatical than, for example, "Golf plays

[14] N. Chomsky, *Syntactic Structures* (The Hague, 1957), p. 15.

[15] It need hardly be said that I have touched here on difficult, even notorious, problems which go beyond the scope of the present investigation. In what follows I shall make no claim to have presented a theory of semantic categories (and thus to have given *clear* sense to "semantical incorrectness"). I grant that on this point, as on many others, the solution I propose depends on the mere promise of a theory. I shall be relying on the possibility of such a theory, and on certain results which, it would seem, any satisfactory theory would yield. See esp. Fred Sommers, "Types and Ontology," *Philosophical Review*, LXXII (1963), 327-363.

sleeps."[16] He speaks of adding to the usual grammatical categories (such as adjective, noun, and so forth) a number of grammatical subcategories (such as animate-noun):

> These are simply a refinement of familiar categories. I do not see any fundamental difference between them. No general procedure has ever been offered for isolating such categories as Noun, Adjective, etc., that would not equally well apply to such subcategories as necessary to make finer distinctions.[17]

Whichever position one takes, there is agreement on this: there is an important intuitive distinction to be reckoned with. For it is clear that expressions such as

> Saturday flies.
> I wrapped a nap in a paper bag.

deviate in some way from the regularities of English, and also clear that they do so in a manner which is different from the way expressions such as

> Saturday airplane.
> Nap paper I bag.

deviate from those regularities. And this is the difference on which the division into two parts of the doctrine of linguistic categories, as I have presented it, is based. I choose to say that the expressions of the second group are ungrammatical—that is, nonsentences; those of the first group, while sentences, are semantically incorrect sentences.

It will be useful to indicate somewhat more carefully what I take to be involved in semantical incorrectness. Our examples are of the familiar sort: "Eleven is purple," "Quantum theory is six inches high," "Virtue is triangular," "Saturday eats algebra," and so forth. Notice, first of all, that semantically incorrect sentences are not odd in the same way that "You look lovely now in the moonlight" (spoken in midafternoon) is odd. The oddity of semantical incorrectness is not simply the inappropriateness of a particular sentence spoken on a particular

[16] *Op. cit.*, p. 78. See also his "Degrees of Grammaticalness" in *The Structure of Language*, ed. by J. Fodor and J. Katz (Englewood Cliffs, 1964), pp. 384-389.
[17] "Degrees of Grammaticalness," p. 385.

occasion; a semantically incorrect sentence has a kind of internal inappropriateness, such that it is odd under any circumstances. Semantical incorrectness, we shall say, applies to sentence types.[18]

It would seem that there are presuppositions of informative discourse, which limit the class of subjects once predicates are given, or limit the class of predicates once subjects are given. The presuppositions are the groupings of things (and of words) into natural classes or categories; these groupings account for part of the effectiveness of our communication of information. I have told you something useful when I tell you that the building where we shall meet is not high (tall), because it is presupposed in such discourse that buildings are either high or low, and so forth. By contrast, there is nothing of importance communicated in telling you that the building where we shall meet is not equal to the square root of 7.

To this it might be objected that metaphors aid in the communication of information precisely through ignoring the boundaries of semantic categories. But I would counter that a notion of semantical incorrectness seems essential to a satisfactory theory of metaphor. Before one can understand an attribution as metaphorical it is necessary that one stop regarding the phrase as functioning merely in the most ordinary way. Beardsley speaks of "discourse that says more than it states, by cancelling out the primary meaning to make room for secondary meaning."[19] Often it is the semantical incorrectness of an attribution that leads us to look for a nonliteral reading.

III. Semantically Incorrect Sentences Are neither True nor False

It is one thing to label a recognizable sort of deviance as semantical incorrectness; it is something else to claim that

[18] Semantical correctness is understood to be relative to the language at a given stage in its history; of course meanings of words change, and thus admit juxtapositions with words of different semantical categories. There are familiar cases of this when phrases change from being metaphorical (when the expressions are, strictly speaking, semantically incorrect) to being "literal" (semantically correct).

[19] *Aesthetic: Problems in the Philosophy of Criticism* (New York, 1958), p. 142.

sentences of this sort are neither true nor false. I wish to defend the latter claim; I shall argue that either we should regard semantically incorrect sentences as neither true nor false, or (what, for our purpose, comes to the same thing) regard semantically incorrect sentences as false, but in a special sense of "false."

The view that such sentences are neither true nor false is common, and the usual argument for it is simple. "Eleven is purple" is surely not true. But, then, is it false? (To answer affirmatively *just* because the sentence is not true would be begging the question—assuming that all sentences are either true or false.) If it is false, then, one would imagine, its denial is true; but are we willing to accept "Eleven is not purple" (or "It is not the case that eleven is purple") as true? The conceptual context in which an ordinary denial—say, "John is not stupid"—is a useful piece of information is that in which we are prepared to accept a disjunction of predicates—family members of some natural class of predicates—as predicable of John. He is bright or stupid or dull or brilliant, and so forth; at least one disjunct holds. On this model, if we accept "Eleven is not purple" as true, we are disposed to accept *some* alternative such as "Eleven is yellow" or "Eleven is red" as true. The sentence seems better regarded as absurd than as true or false. Since "Numbers are colored" is false, "Eleven is purple" is neither true nor false. Why should we strain to decide upon its truth value when to decide in either direction is unsatisfactory or misleading?

To this familiar argument the following objection may surely be raised. Though "Eleven is purple" may not be false in precisely the same way as, say, "Lettuce is purple" is false, still it *is* false; for the fact is that the number eleven is *not* a member of the class of purple things. Arthur Pap has suggested a way of admitting this objection while still maintaining the important intuitive distinction between (ordinary) false sentences and semantically incorrect sentences.[20] We could call sentences such as "Eleven is

[20] "Types and Meaninglessness," *Mind*, 69 (1960), 41-54. Pap's distinction is between ordinary falsity and meaninglessness (where by "meaningless" he seems to mean "neither true nor false").

purple," "Virtue is triangular," and so forth not absurd, but rather *unlimitedly false*. Sentences like "The basketball is triangular" or "Lettuce is purple" would be called *limitedly false* (that is, false in the ordinary sense of "false"). The point is that to say a given sentence is limitedly false is to say that its predicate is not truly predicable of what is denoted by its subject, and that some other predicate of the same category as its predicate *is* truly predicable of what is denoted by the subject[21]; to say that a sentence is unlimitedly false is to say that its own predicate is not truly predicable of what is denoted by its subject, but also that *no* predicate of that category is truly predicable of what is denoted by the subject. "Lettuce is purple" is limitedly false because in denying it one can agree that *some* color-word (that is, a word belonging to the same semantical category as "purple") applies to lettuce. "Eleven is purple" is unlimitedly false because eleven is not only not purple, it has no color at all—it is not the sort of thing that has a color.

Pap's treatment of the matter has several advantages over simply denying that truth and falsity apply to semantically incorrect sentences. First of all, it focuses attention on our reason for denying that semantically incorrect sentences are true or false (in the ordinary sense)—that not just one predicate, but rather a whole (natural) class of predicates fail to be truly predicable of whatever is denoted by the subject. Secondly, it allows us to avoid the unclear label "absurd," giving us instead the relatively clear "unlimitedly false." But either way of looking at the matter—semantically incorrect sentences as neither true nor false, or as unlimitedly false— will suffice for our purposes. For suppose I have shown that the sentence (a) is semantically incorrect. It will stop the deduction of the contradiction from (a) to have agreed either (i) that semantically incorrect sentences are neither true nor false or (ii) that semantically incorrect sentences are unlimitedly false. For in the first case the law of excluded middle, needed for the deduction, is seen to be false. Only the qualified version of it—"every seman-

[21] This is a very rough characterization; it will not do, for example, for "It is raining."

tically correct sentence is true or false"—will be allowed, and in this form it does not apply to (a) which is semantically incorrect. In the second case, we can conclude from the fact that (a) is unlimitedly false that (a) is neither true nor false. That is, not only is (a)'s own predicate, "is false," not truly predicable of (a) —the denotation of (a)'s subject expression—but the category mate predicate, "is true," is also not truly predicable of (a). The situation then is the same as that of the first case.

It would suffice, then, to show that (a) is semantically incorrect.[22]

IV. Procedure for Determining Semantical Correctness

In this section I shall sketch a general procedure for distinguishing semantically correct from semantically incorrect sentences, confining my attention to simple (noncompound) subject-predicate sentences. In the next section I shall argue that this is an intuitively sound procedure, that is, that it captures reasonably well the *explicandum* of semantical correctness. Then, in Section VI, I will show that the procedure bars the paradoxical sentence (a) from being judged semantically correct.

The most important methodological feature of the present investigation is this: I take it to be crucial that in justifying the procedure outlined below, no reference will be made to the success of the procedure in barring (a). The justification of the procedure will rest entirely on the question of whether or not the procedure provides a good explication of the notion of semantical correctness. It is precisely in this sense that the proposed solution to the Liar is not *ad hoc*; it is of course irrelevant that one's *motive* for setting up the procedure is to stop the contradiction. A solution to the Liar, as opposed to a removal of the contradiction (a matter of no difficulty at all), consists in showing

[22] Actually, I shall argue only for the weaker conclusion: that it is impossible to show that (a) is semantically correct (see Section VI). This weaker conclusion seems sufficient, however, to stop the deduction of the contradiction.

that one's way of removing the contradiction is natural and intuitive.[23]

I will try to show that the way offered here is natural and intuitive, but it is no doubt the case that this will *need* to be done; the procedure is not obviously, or at first sight, natural (or the Liar would not be as serious a puzzle as it is). For it should be clear at the outset that (a) is not obviously semantically incorrect.

A. The first step in determining whether a given sentence is semantically correct is to *determine the range of applicability of the predicate of the sentence.* Actually to specify the kind of objects to which a given predicate may be applied (so that, as a result, the sentences are true or false) may often be a task of great difficulty. The ranges of applicability of predicates are frequently changed in the life of a natural language, and, as Austin has pointed out, there are simply no prearranged determinations for our use of our ordinary expressions in extremely unusual situations. However, the intuitive basis of determining a range of explicability (that is, what it *is* to do so) seems relatively clear. If one knows how to use a particular predicate successfully, then one knows, a fortiori, at least roughly to what sort or sorts of things the predicate applies. And conversely, one learns how to use an unfamiliar predicate partly by learning to what sorts of things the predicate applies; for example, part of distinguishing two senses of the word "radical" is learning that in one sense it applies to numbers, in another to humans.

For our purposes it will not be necessary that one be able to specify ranges of applicability very precisely (and it is surely wrong to suppose that one does not understand a predicate unless

[23] This point seems to have been overlooked in L. Goddard's article, "Sense and Nonsense," *Mind*, 73 (1964), 309-331, where he understands the argument to a contradiction (in any self-reference paradox) as a simple *reductio*: "there is no contradiction to be avoided but simply a straight conclusion to be accepted" (p. 328). His argument seems to be that the contradiction shows that one of the premises—an existential assumption—is false. It is true that the contradiction shows that *some* premise is false, but it is precisely because each of the premises is plausible to some degree that we have paradox and not *reductio*. The undertaking called for is just what Goddard opposes; it is "to use the contradiction as a motive for creating a theory" (p. 328) with independent support and of sufficient strength to show that some premise of the argument leading to contradiction lacks plausibility after all.

one *can* specify its range of applicability precisely); in the cases which will be of particular interest there should be no special problem in reaching agreement.

B. The second step of our procedure is to *determine whether or not the sentence is self-referential.* Let us say that a sentence type is self-referential if and only if every token of that type is self-referential, and that a sentence token is self-referential if and only if what one *mentions* with the subject expression of the sentence token is the sentence itself (type or token) which is being *used.* Sentences beginning with the words "This very sentence" or "This very expression," for example, are self-referential sentence types; "The next sentence is true" has, as we have seen, some self-referential tokens.

C. The third step of our procedure applies only to non-self-referential sentences. It tells us to *determine whether or not the subject of the sentence indicates, on the basis of its sense, the sort of thing included in the range of applicability of the predicate; and to call the sentence "semantically correct" if and only if it does.*

I distinguish here between the sense and reference of an expression after Frege's distinction between *Sinn* and *Bedeutung.*[24] The expression "The brownstone on the corner of 56th and Broadway" has reference if and only if there *is* a brownstone on the corner of 56th and Broadway. But it has sense even if there is no such edifice; I know what you mean when you say, "The brownstone on the corner of 56th and Broadway is for sale," even if, in fact, there is no such building. The distinction between sense and reference that I wish to draw is different from Frege's in this respect: I wish to speak of the sense of an expression as indicating a certain sort or kind of thing and, as far as I know, Frege did not speak this way. Suppose someone says, "The brownstone on the corner of 56th and Broadway is for sale," and in fact there is no such building. Although the subject of that sentence has no reference, it seems reasonable to ask, "What *sort* of thing is that sentence about?" and reasonable to reply, on the basis of the sense of the expression, "A house—that sort

[24] See "On Sense and Reference," *Translations from the Philosophical Writings of Gottlob Frege,* by P. Geach and M. Black (Oxford, 1960), pp. 160-178.

of thing," or "A brownstone house—that sort of thing," or even, "A physical object—that sort of thing." Such an answer does not commit one to the existence of the house referred to; but it does enable us to apply our test for semantical correctness to sentences whose subjects lack reference.

D. The fourth step of our procedure applies only to self-referential sentence types. (Sentences, some of whose tokens are self-referential and some not, are not treated by either step C or D of this procedure; for such sentences it is token oddity and freedom from token oddity which are the relevant notions.) We are directed, for all self-referential sentence types, to *determine whether the demonstrative reference of the subject is included in the range of applicability of the predicate, and to call the sentence "semantically correct" if and only if it is.*

I distinguish here between the demonstrative reference and the reference proper of an expression. This terminology is designed to enable us to speak with clarity about the familiar situations in which expressions misdescribe what they nevertheless manage to refer to (or, if one prefers, situations in which we misdescribe what we nevertheless manage to refer to).[25] Let us say that an expression may have demonstrative reference as well as reference proper if and only if, in distinguishable ways, it both indicates (identifies) an object and also describes or characterizes that object. Typical cases involve the use of the demonstratives "this," "that," and so forth, in certain situations; but it is possible for expressions without demonstratives to have both sorts of reference. "The man in the distance" has reference proper just in case there is a man in the distance; its demonstrative reference is whatever, in a particular situation, is referred to—whatever is spotted in the distance. Similarly, "this very Latin expression" has demonstrative reference, but no reference proper.

We find shortcuts in producing descriptions used to identify objects. We may be spared the trouble of saying, "The copy of *Moby Dick* John loaned me is full of marks" by having the book close at hand; we say simply, "That copy of *Moby Dick* is full

[25] See K. S. Donnellan's analysis of such situations in "Reference and Definite Descriptions," *Philosophical Review*, LXXV (1966), 281-304. Unfortunately I learned of this work too late to bring it to bear on the present discussion.

of marks." But the price we pay for this convenient mixing of demonstrative and descriptive identification is the possibility of a discrepancy between them; the book I am pointing to may in fact be *The Brothers Karamazov*. The expression "that copy of *Moby Dick*" cannot have reference proper if it lacks demonstrative reference, but it can have the latter without (and also, of course, with) the former. It should be emphasized that this extension of Fregean terminology does not distort the familiar distinction of Frege's between sense and reference; it only gives a name to an object which is misdescribed by an expression that still refers to it. This sort of misdescription plays a crucial role, I shall argue, in the crux version of the Liar Paradox.

V. Justification of the Procedure

There seem to be two features of the procedure described above that need justification. The first is that the procedure does not apply to sentences *some* of whose tokens are self-referential; it applies only to non-self-referential sentences (sentences with no self-referential tokens) and to self-referential sentence types. I will try to show, when I turn to other versions of the Liar, that it is more profitable to apply a specially defined notion of token oddity to sentences excluded from consideration here. But the theoretical question remains: how can we know that all (or no) tokens of a given sentence type are self-referential? Unless there is, in principle, some way of isolating the non-self-referential and the self-referential sentence types, there would seem to be no way of knowing when to apply the procedure, for it applies to just these.

To begin with, the question of whether or not a given sentence type has self-referential tokens is an empirical question. Just because we know of no ordinary, or even extraordinary, context in which a token of a particular sentence would be seen to refer to itself, there is no certainty that there *is* no such context. And just because a given sentence is such that every situation we can imagine involves the sentence in self-reference, there is no guarantee that the sentence *is* a self-referential sentence type. But of course our procedure can tell us that *if* the given

sentence is non-self-referential, then it is semantically correct, and that *if* a given sentence is indeed a self-referential sentence type, then it is semantically correct. It may be reasonable to assume, until further notice, that a particular sentence satisfies one of the two antecedents; then it is reasonable to assume, until further notice, that the sentence is as the procedure tells us. It is possible that we shall find that we were wrong in applying the semantical-correctness procedure, and find that we should raise, instead, the question of token oddity.

The second feature of the procedure which calls for justification is the separating of self-referential and non-self-referential sentences for different treatment. For it is not obvious, first of all, that the question of semantical correctness has anything to do with the distinction between self-referential and non-self-referential sentences or, secondly, that the question of semantical correctness has anything to do with the distinction between sense and demonstrative reference. Notice that the procedure advocated does not commit us to the view that the semantical correctness of a non-self-referential sentence is something different from the semantical correctness of a self-referential sentence, but it does commit us explictly to the view that the criteria for determining the semantical correctness of the two kinds of sentences are different. This view calls for justification. I shall assume that the procedure for determining the semantical correctness of non-self-referential sentences (step C of the procedure) is acceptable, and consider only, in what follows, the question: why should we determine the semantical correctness of a self-referential sentence on the basis of the demonstrative reference of its subject expression (its inclusion in the range of applicability of the predicate) instead of, as with non-self-referential sentences, determining semantical correctness on the basis of the *sense* of the subject expression?

My argument in support of this policy begins with the simple observation that, when confronted with referring expressions that combine the two sorts of reference (such as "that porcelain cup"), one does not assume that, because *something* is indicated (the expression has demonstrative reference), the thing indicated is as described (the expression has reference proper). The demon-

strative reference of "that porcelain cup" may indeed be made of plastic. What is called for, if one cares whether the expression has reference proper, is an examination of the demonstrative reference.

The reason one tests the semantical correctness of ordinary (non-self-referential) sentences on the basis of the *sense of* the subject is, I take it, that the subject may not *have* reference of any sort, and yet it may still be semantically correct. A procedure for testing the semantical correctness of such sentences as "The present king of France is bald" must, it seems obvious, rest on the sense, not the reference (of any sort), of the expression "the present king of France."

But when we test the semantical correctness of self-referential sentences, it should be clear, first of all, that we do not *need* to base our inquiry on the sense of the subject. For with self-referential sentences, we have before us, open to inspection, the thing indicated by the subject of the sentence. For such sentences there is no possibility that the subject lack both kinds of reference; by hypothesis, all self-referential sentence types have subjects which refer to the very expressions of which they are part. For any such sentence, the reference (the demonstrative reference) of the subject is precisely the expression with which we are dealing. Hence the crucial question concerning semantical correctness —does the subject indicate the something of the sort included in the range of applicability of the predicate?—*may*, in the case of self-referential sentences, be posed in terms of the demonstrative reference of the subject; specifically, the question may be, does the demonstrative reference of the subject fall within the range of applicability of the predicate? Of course the sort of reference we should consider here is the demonstrative reference; as pointed out above, we avoid assuming because *something* is referred to by an expression (for example, "this very superbly elegant phrase") that that something is as described.

But not only *may* we pose our inquiry into semantical correctness on the basis of the demonstrative reference, not the sense, of the subject of a self-referential sentence; we *ought* to do this. For if we grant that there is no special reason to deal with the sense instead of the demonstrative reference of subjects of self-referential

sentences (granted, I am assuming that the only possible reason was reference failure), it would seem clearly preferable to deal with the demonstrative reference in our determination of semantical correctness. For the demonstrative reference is immediately before us, open to our unprejudiced inspection as to its various syntactical and semantical properties. The sense of the subject, on the other hand, indicates reference, as Frege said, and it seems clear that the reference which the sense indicates is the reference proper.

The objection might be raised: semantical correctness should be, properly speaking, a question of meanings (that is, senses), not a question of objects referred to. But this objection, so far as I can see, begs the question. We say this because, in the cases of the sentences with which we mostly deal (non-self-referential sentences) we need to consult meanings or senses of terms, instead of references, in order to treat cases involving subjects without reference. Our experience with sentences of this sort is so predominant that we can easily be misled into thinking that one must always regard semantical correctness in this way.[26] But there are sentences where there is no danger of the subject lacking reference, and for these there seems to be no reason to have recourse to the senses of the subjects.

VI. Application of the Procedure to the Sentence (a)

Assuming now that we have before us a roughly satisfactory explication for the notion of semantical correctness, in the form of a procedure for testing simple subject-predicate sentences of

[26] This, I should say, is precisely the error on account of which the Liar seems so incorrigible. We presume that semantical correctness, for all sentences, must be decided on the basis of meanings or senses of terms; then, of course, "This very (semantically correct) sentence is false" is semantically correct, since its subject indicates, by its sense, the sort of thing that *can* indeed be false. Further, we feel that by saying only that the subject indicates the *sort of thing* that can be false, we are uncommitted as to whether the subject really has reference proper. But of course once we conclude that the sentence is semantically correct, it follows that the subject has reference proper. From an apparently innocent start we find ourselves once again confronted with paradox, for semantically correct sentences are true or false, and this one, it can be shown, is both true and false.

English, we proceed to an application of this procedure to the sentence:

(a) This very sentence is false.

One preliminary point: I shall construe (a) as:

This very (semantically correct) sentence is false,

for ordinarily one speaks of semantically correct sentences simply as "sentences." For example, when one says "Only the first sentence of his article is interesting," one would not be taken to mean that the sentence in question is *merely* grammatical.[27]

In accordance with the first step of the procedure we determine the range of applicability of the predicate "is false." I have already urged that we accept the characterization of the class of semantically correct sentences as those sentences which are either true or false. It seems equivalent to this characterization to state that the range of applicability of the predicate "is false" and its category mate "is true" is the class of semantically correct sentences. I take it that " 'Snow is white' is false" is false but semantically correct, while " 'Saturday eats algebra' is false" is neither true nor false, but rather semantically incorrect. It is semantically correct to call a sentence true or false just in case the sentence is semantically correct.

Taking (a), as above, as a self-referential sentence type, we next ask, according to the procedure, whether the demonstrative reference of its subject is included in the range of applicability of its predicate, "is false." But the demonstrative reference of the subject of (a) is (a) itself; hence, before we can proceed, we must determine whether (a) is semantically correct. Here of course our inquiry breaks down; (a) can pass the test devised for self-referential sentences only after it passes that test. Granted, the procedure does not yield the result that (a) is semantically incorrect; what we see, rather, is that (a) can never reasonably be ruled semantically correct. Semantical correctness, then, though it is not an effective notion, is strong enough to rid us

[27] Of course, even if one insists that (a) be understood as concerning a mere sentence, it follows from the procedure that (a) cannot be judged semantically correct.

of the plausibility of the argument of the Liar Paradox. Since (a) can never be shown to be semantically correct, it is unreasonable to apply the law of excluded middle to (a)—that is, one has grounds for denying step (3) of the argument of the paradox.

VII. Extension of the Proposed Solution

Earlier I distinguished three disguises for the self-reference involved in the Liar. That is, sometimes the self-reference is indirect, sometimes empirical, and sometimes imbedded in generality of reference. Versions of the Liar with these sorts of self-reference differ from the crux version, as I have pointed out, in involving sentences only some of whose tokens are self-referential. It is clear that a different sort of solution is called for here: the sentences involved can function, in some circumstances, in perfectly normal ways. There is certainly nothing wrong, in *general*, with, for example, "The next sentence is true" or "The previous sentence is false" or "Sentence 13 is false" or "Everything said by a Cretan is false."

Let us give the name "token oddity" to the kind of oddity a sentence token may have, owing to the semantical incorrectness of some sentence type. Let us say, for example, that a token of "That is triangular," uttered in a situation in which "that" clearly refers to, say, impatience, is token odd. Such a token of "That is triangular" is token odd because the sentence type "Impatience is triangular" is semantically incorrect. The intuitive point of this terminology is simply that a situation or context can bring to a particular sentence token what is secured permanently (for every token) for some other sentence. The sentence type "Impatience is triangular" is semantically incorrect because the sense of its subject expression indicates something of the sort that does not fall within the range of applicability of its predicate. A token of "That is triangular" is token odd, then, just in case the situation causes "that" to refer to something that does not fall within the range of applicability of "is triangular." Let us say, with Quine, that truth and falsity are traits of sentence

types at particular times (on particular occasions).[28] I have already argued that semantically incorrect sentences are neither true nor false; token oddity is the same characteristic as it applies to sentence tokens. The token-odd occurrence of "That is triangular" is neither true nor false.

Here is the strategy of what follows: I will show that the context involved in the versions of the Liar in question bring to the tokens of sentences involved just those features which, secured permanently in "This very sentence is false," make the latter sentence semantically incorrect. Strictly speaking, since we have not shown "This very sentence is false" semantically incorrect (we have shown only that it can never be shown semantically correct), we cannot show the sentence tokens involved in the other versions of the Liar to be token odd in the same sense as, in the previous example, "That is triangular" is token odd. All we can show is that these sentence tokens cannot be shown to be free of token oddity; hence there is no ground for regarding them as true or false, and no contradiction which depends upon applying excluded middle to them can be deduced.

There is another, closely related, way of viewing the matter. Suppose we can see that because of their context, a group of (one or more) sentences does involve self-reference. Then let us apply to each sentence token of the group that part of the procedure for testing semantical correctness of sentence types that we should apply if its every token were self-referential. Remember that this will, at best, establish only token oddity and not semantical incorrectness; this may be regarded as a hypothetical application of the procedure.

Consider a typical paradoxical case of indirect self-reference:

(1) The following semantically correct sentence is true.
(2) The following semantically correct sentence is true.
(3) The semantically correct sentence before last is false.

[28] *Word and Object* (Boston, 1960), p. 191. Cf. Aristotle: "The same statement, it is agreed, can be both true and false. For if the statement 'he is sitting' is true, yet, when the person in question has risen, the same statement will be false" (*Cat.* 4a23).

(Once again I construe an ordinary occurrence of "sentence" as "semantically correct sentence.") The juxtaposition of these sentences involves the tokens in (indirect) self-reference. This is one of the features present in the sentence type "This very sentence is false." The other feature is present also: in both cases there are predicates whose ranges of applicability is limited to semantically correct sentences. Since, then, the sentence tokens under consideration have, partly through context, the features that are secured permanently in "This very sentence is false," it follows that they should be regarded as neither true nor false. This stops the deduction of contradiction from them.

Here is a hypothetical application of the procedure. Since the range of applicability of the predicate of (1) is limited to semantically correct sentences, we see that it is semantically correct if and only if the demonstrative reference of its subject expression is semantically correct. That is, (1) is semantically correct if and only if (2) is. But, noting that the predicate of (2) is also "is true," we see that (2) is semantically correct if and only if its demonstrative reference (3) is semantically correct. Similarly, (3) is semantically correct if and only if (1) is. Here, again, our inquiry breaks down; we can know that (1) is semantically correct, it turns out, only by finding out whether (1) is semantically correct; that is, we can never know. But our application of the procedure is only hypothetical; since it is only a token of (1) that is self-referential, it is only a token of (1) that cannot be shown free of token oddity.

Consider a typical paradoxical case of empirical self-reference:

The semantically correct sentence on line 28 is false.

Again, the sentence type involved is not self-referential; it is only the token of this type that occurs at line 28 that is self-referential. But this token of the sentence has the two features which cause it to be impossible to judge "This very sentence is false" semantically correct: (*i*) it *does* refer to itself; (*ii*) it does have a predicate whose range of applicability is limited to semantically correct sentences. Therefore, the displayed token cannot be shown free of token oddity. The same thing may be shown by a hypothetical application of the procedure as above.

Some versions of the Liar involving generality of reference make use of sentences only some of whose tokens are self-referential. But let us deal first with such versions involving sentence *types*. How does the procedure apply here? Consider the sentence "Every sentence is false," which, as before, I shall construe as:

(G) Every semantically correct sentence is false.

There is, of course, the question of whether the subject of this sentence has reference of either sort: there does not seem to be one thing (a class or anything else) that it names. I doubt, however, whether that issue bears crucially on the problem before us. It seems legitimate to say that (G), however it is to be characterized, is the demonstrative reference of the subject of (G). This is a way of expressing the intuitively clear fact that a sentence about every sentence is about itself as well. If indeed (G) is the demonstrative reference of the subject of (G), then the semantical correctness of (G) depends upon whether (G) is semantically correct; once again we find our procedure cycling helplessly, showing us that there can be no sufficient reason for calling (G) semantically correct, and thus stopping the derivation of a contradiction from (G).

We have next to consider cases where sentence tokens are self-referential (again, through generality of reference). The sentence

(C) Everything said by a Cretan is false [said by a Cretan]

will serve. Here, as in the other cases, such an utterance has the features secured permanently in "This very sentence is false" —self-reference, and a predicate whose range of applicability is limited to the class of semantically correct sentences. Hence, (C) is token odd; a hypothetical application of our procedure—it would amount to what we have just done concerning (G)—yields the same result.

VIII. The Proposed Solution Does Not Rule Out All Self-Reference

It remains to show, as claimed initially, that this solution to the Liar has the advantage of distinguishing between two classes

of self-referential sentences, the members of one of which do pass the test of semantical correctness. Consider for example:

(1) This very sentence is in English.

The range of applicability of "is in English" would seem to be limited only to the class of linguistic entities; surely semantically incorrect sentences, even mere expressions, can be said, truly or falsely, to be in English. Since (1) is self-referential, we apply the final step of the procedure, asking whether the demonstrative reference of the subject of (1) is included within the range of applicability of the predicate. Now whether or not the subject of (1) is construed, as earlier, as "this very (semantically correct) sentence," the demonstrative reference of the subject is surely a linguistic entity. Here we are not forced to ask the question-begging question: is (1) semantically correct? The predicate "is in English" is, as it were, less demanding. Therefore, (1) is semantically correct.

It should be clear that the procedure applies with the same result to

(2) This very sentence is interesting

and to

(3) Everything I say, including what I am now saying, is carefully considered

since the predicates of both are applicable to the sorts of things which the subjects of (1) and (3) have as their demonstrative references.

Even such a sentence as

(4) This very sentence is semantically correct

passes our test. For the range of applicability of "is semantically correct" surely includes, among other things, merely grammatical sentences; it is semantically correct, though false, to say that "Impatience is orange" is semantically correct. And it is apparent upon inspection that (4), the demonstrative reference of the subject of (4), is a grammatical sentence.

What sentences, besides "This very sentence is false" and "This very sentence is true," are barred by our procedure? "This very sentence is dubious" or "This very sentence levels a serious

accusation," for example, would seem to be in the same position as the paradoxical sentence; they cannot be judged semantically correct, for their predicates seem applicable only to semantically correct sentences. We do not ordinarily say or deny that mere expressions, or merely grammatical sentences, are dubious, or level accusations; it would seem to require the full capabilities of the semantically correct sentence to be or not be dubious, to level serious or mild accusations.

It may be useful to mention here, with one example for each, some of those sentences called to mind by versions of the Liar other than the crux version, which can now be sanctioned. Starting with indirect self-reference, consider the three sentences:

(6) The following semantically correct sentence is about the one which follows it.

(7) The following semantically correct sentence is about the first of this group.

(8) The first semantically correct sentence is about the one which follows it.

It is clear that indirect self-reference is involved, owing to the juxtaposition of tokens of the three sentences; hence each sentence token has one of the features of the ill-fated sentence, "This very sentence is false." But it is not at all clear that the other feature—the presence of a predicate whose range of applicability is limited to semantically correct sentences—is present in this group of sentence tokens. For it would seem that there is a sense of "about" in which a mere expression could be said, truly or falsely, to be about something else. Indeed, on the basis of a hypothetical application of our procedure, it is clear that the sentence tokens involved are free from token oddity. Begin with (6), which is, in this context, self-referential. (6) is semantically correct (that is, if it were a self-referential sentence *type*, it would be semantically correct) just in case the demonstrative reference of its subject is an expression; (7) is the demonstrative reference of the subject of (6), and of course it is an expression, so (6) is semantically correct (that is, is free of token oddity, since the application of the procedure is hypothetical). In the

same way, one can determine that (7) and (8) are free of token oddity.

Consider the sentence:

(9) Sentence 9 contains ten words.

Here, by the empirical accident of having the number 9, a token of "Sentence 9 contains ten words" becomes self-referential. But there is no token oddity here, for it is clear by inspection that the demonstrative reference of the subject of that token (whether the demonstrative reference is understood to be the token or the type) is at least an expression composed of words, and as such does fall within the range of applicability of the predicate "contains ten words."

Notice, finally, that any self-referential token of "Every Cretan sentence is in Greek" (one uttered by a Cretan) is free of token oddity.

IX. Sentences with Complex Predicates

I shall consider, as an example, only one complex predicate, "is true or false."[29] There would seem to be some difficulty in treating:

(10) This very sentence is true or false.

The difficulty may be put in the following way. A sentence may, we have seen, say of itself that it is semantically correct; a sentence is semantically correct if and only if it is either true or false; however, a sentence may not, we have seen, say of itself either that it is true or that it is false.

[29] By means of the example, "This very sentence is either false or *UF*" (see n. 21), the editors have made me see a further restriction that would be needed in extending the present solution to handle complex predicates. The restriction is (roughly): no compound predicate applies to more than what its least applicable component applies to. For example, the range of applicability of "is red and colored" is no more inclusive than that of "is red." That contradiction arises without this restriction can easily be seen with the simpler example, "This very sentence is false or in French" ("This very sentence is false and in English" would serve as well). With the restriction, it can be shown, according to the procedure sketched above, that the sentence can never be shown to be semantically correct.

I propose that we agree to call (10) semantically correct, since its disjunctive predicate exhausts a particular category. I take it, that is, that "is true or false" covers the same category covered by the single predicate "is semantically correct," just as "is 1 or 2 or . . . and so forth" covers the category covered by "is a positive integer." I propose that we understand a disjunctive predicate that exhausts a whole category as having the range of applicability of the single predicate which is the family name of that category.

All that is required, if we treat the matter this way, is that we deny that in general the semantical correctness of a sentence of the form "x is P" follows from the semantical correctness of a sentence of the form "x is P or Q." In particular, if "P or Q" exhausts a category whose individual members do not include in their ranges of applicability the demonstrative reference of whatever replaces x in the sentence form, the inference from the semantical correctness of "x is P or Q" to the semantical correctness of "x is P" or "x is Q" is invalid.

As a consequence of allowing (10) as semantically correct, it follows that nothing stands in the way of showing that "Every sentence is true or false" (or, in its properly restricted version, "Every semantically correct sentence is true or false") is semantically correct. Thus the present solution enables us to formulate the law of excluded middle in what is surely the most natural way, and in a way which is not open to adherents of the language-level approach to the semantical paradoxes.[30]

X. Treatment of One of Quine's Versions of the Liar

Professor Quine has provided us with a particularly interesting verion of the Liar;[31] actually, this version can be understood in two different ways, depending on where one takes the subject and predicate to be divided.

[30] On this and other philosophical disadvantages of the language-level approach see F. B. Fitch, "Universal Metalanguages for Philosophy," *Review of Metaphysics*, 67 (1964), 396-402, and "Self-Reference in Philosophy," *Mind*, 55 (1946), 64-73, rev. and reprinted as Appendix C of Fitch, *Symbolic Logic* (New York, 1952), pp. 217-225.

[31] See, e.g., *Set Theory and Its Logic*, p. 255.

(Q1) "Appended to its own quotation is a falsehood" appended to its own quotation—is a falsehood.

(Q2) "Becomes part of a falsehood when appended to its own quotation"—becomes part of a falsehood when appended to its own quotation.

Following the usual convention, let us regard an expression with quotation marks to the left and right of it as a *name* of the expression appearing between the quotation marks. It should be noted first that only (Q1) is self-referential in the narrow sense we have adopted: that is, the subject of (Q1)—the part before the dash—does refer to the whole sentence (Q1), whereas the subject of (Q2) refers to (names) only the expression quoted.

Consider the self-referential (Q1) first. An application of our procedure succeeds in stopping the derivation of a contradiction from (Q1), because it shows that (Q1) cannot reasonably be regarded as semantically correct, and hence true or false. For to find (Q1) semantically correct, we should have to know that the demonstrative reference of its subject (and the demonstrative reference *is* [Q1]—another token of the type [Q1]) is semantically correct; this is the point at which the procedure cycles.

But the situation with (Q2) is more problematic. Since the sentence is not self-referential, we test for the semantical correctness of it on the basis of the *sense* of its subject expression. It is doubtful whether one should speak of a quotation-name as having a sense; the object named is actually presented, between the quotation marks. There is no possibility that a quotation-name should fail to name (fail to have reference). At least it is clear, through examples, what it would be like for a sentence with a quotation-name as subject to be semantically incorrect. " 'Roses are red' is less than 4" is semantically incorrect because sentences are not numbers. We base our judgment here on whether the expression actually present (between the quotation marks) is included in the range of applicability of the predicate.

What can be learned of the expression

becomes part of a falsehood when appended to its own quotation

—the expression named by the subject of (Q2)—by inspection? We can see only that it is an expression; we certainly cannot assume that it becomes part of a semantically correct sentence when it is appended to its own quotation (producing another token of [Q2]). That is the question before us: is (Q2) semantically correct?

Now what is the range of applicability of the predicate "becomes part of a falsehood when appended to its own quotation"? I suggest that it is the class of expressions which become parts of semantically correct sentences when they are appended to their own quotations. It can be determined, for example, that the expression "is in English" is such that it becomes part of a semantically correct sentence when it is appended to its own quotation; and this without begging any questions. That is,

"is in English" is in English

is semantically correct just in case what is named by its subject is the sort of thing that can be in English, or fail to be in English. And what is named by the subject *is* that sort of thing, so the sentence is semantically correct. *Now* we know that the expression "is in English" is something of which it can be said (truly or falsely—in this case falsely) that it becomes part of a *falsehood* when appended to its own quotation. Consider, for contrast, the expression "puzzles are fun"; this, it seems to me, falls outside the range of applicability of the predicate, "becomes part of a falsehood when appended to its own quotation." That is to say, I regard

> "puzzles are fun" becomes part of a falsehood when appended to its own quotation

as semantically incorrect. That expression is not the kind of expression that becomes part of a truth *or* falsehood when it is appended to its own quotation because it is not the kind of expression that becomes part of a semantically correct sentence when it is appended to its own quotation.

The point is that the predicate in question demands (for semantical correctness) not only that the object of which it is predicated be an expression, and not only that when it is appended to its

quotation it be a sentence, but also that the sentence so formed be a semantically correct sentence; what the predicate leaves open (as far as using it in a semantically correct sentence is concerned) is whether the sentence so formed is true or false, for on that matter it *says* (rather than presupposes) something.

So the test of the semantical correctness of (Q2) depends upon whether the expression named by the subject of (Q2) is such that it becomes part of a semantically correct sentence when it is appended to its own quotation. That is, it depends upon whether the following sentence is semantically correct:

> "becomes part of a falsehood when appended to its own quotation" becomes part of a falsehood when appended to its own quotation.

But the displayed sentence is (Q2) itself, so our inquiry is frustrated. Since it is impossible to establish the semantical correctness of (Q2), there is no reason to regard it as true or false; this stops the derivation of a contradiction from (Q2).

XI. Summary

My principal motivation for developing the present proposal for solving the Liar was the feeling that the solution that follows simply from the language-level theory is not adequate. A solution to the Liar does follow from the language-level theory only by understanding it to involve not only the claim that there *are* levels of language, but also the claim that no sentence may appear at two levels at once (that is, that no sentence may mention itself). It seemed advisable to give up the language-level theory's assumption about the distinctness of the levels of language, and investigate the problem from a different starting point, that of linguistic categories and semantical correctness. The proposed solution is not intended as evidence against the existence of levels of language, but only as a better treatment of the problem of self-referential sentences than that which follows simply from the language-level theory. The present solution would in no way

commit one to the rejection of the use-mention distinction on which the language-level theory is based; it only contradicts the language-level theory's unqualified restriction against the coincidence of use and mention, by sanctioning many self-referential sentences.

ROBERT L. MARTIN

State University of New York, Buffalo

PRESUPPOSITION, IMPLICATION, AND SELF-REFERENCE *

by BAS C. VAN FRAASSEN

THE two aims of this paper are, first, to explicate the semantic relation of *presupposition* among sentences, and, second, to employ the distinctions made in this explication in a discussion of certain paradoxes of self-reference. Section I will explore informally the distinction between presupposition and im-

[Continued on next page]

* An. earlier version of this paper was read at Duke University on May 12, 1 7, and se tions I and II were included in a paper presented at a symposium on free logic held at Michigan State University on June 9 and 10, 1967. Acknowledgments and bibliographical references have been collected in a note at the end.

plication. In section II we construct a formal paradigm in which this distinction corresponds to a substantial difference. Section III explores the relations between presupposition and truth, and, in section IV, the *Liar* and some similar paradoxes are finally broached.

I

The best known source for the concept of presupposition is the view (a.o. Strawson's) that a property cannot be either truly or falsely attributed to what does not exist. Thus, the sentence "The King of France (in 1967) is bald" is neither true nor false, on this view, *because* the King of France does not exist.[1] The explicit characterization of *presupposes* is therefore given by

1. *A presupposes B* if and only if *A* is neither true nor false unless *B* is true.

This is equivalent to

2. *A* presupposes *B* if and only if
   (a) if *A* is true then *B* is true,
   (b) if *A* is false then *B* is true.

From this equivalence it is clear that presupposition is a trivial semantic relation if we hold to the principle of bivalence (that every sentence is, in any possible situation, either true or false). In that case, every sentence presupposes all and only the universally valid sentences.

Why can we not say that, on Strawson's position, "The King of France is bald" *implies* "The King of France exists"? The reason is that presupposition differs in two main ways from implication, on any accepted account of the latter relation. For implication, *modus ponens* is accepted as valid; from 2(a) it is clear that an analogue of *modus ponens* holds also for presupposition. But *modus tollens* is also generally accepted as valid for implication; what of its analogue for presupposition? This question cannot be answered unless we first settle on the meaning of negation. We shall understand the negation of a sentence *A* to be true (respectively, false) if and only if *A* is false (respectively, true). This is not the only possible convention, but it has the virtue of yielding the convenient further characterization of presupposition:

3. *A* presupposes *B* if and only if:
   (a) if *A* is true then *B* is true,
   (b) if (*not-A*) is true then *B* is true.

[1] The question whether a sentence is true may be said to make sense only relative to an interpretation, or to what is intended by it, and so on. Except when dealing with artificial languages, we shall avoid this issue by assuming that there is a "correct" interpretation known to the reader.

From 3 it is clear that the analogue to *modus tollens*

4. *A* presupposes *B*
   (not-B)
   Therefore, (not-*A*)

is not valid; if the premises are true, the conclusion is not true (and not false). Secondly, 3(b) shows that the argument

5. *A* presupposes *B*
   (not-A)
   Therefore, *B*

is valid: if its premises are true, so is its conclusion. But of course the analogue of 5 for implication does not hold.

Thus presupposition and implication are not the same, but they have something in common. What they have in common is that, if *A* either presupposes or implies *B*, the argument from *A* to *B* is valid. This is itself a semantic relation, which we shall here call "necessitation":

6. *A necessitates B* if and only if, whenever *A* is true, *B* is also true.[2]

We see here an obvious possible cause of confusion between implication and presupposition. The standard logic text generally begins with an explanation that an argument is valid if and only if the conjunction of the premises implies the conclusion, if and only if the corresponding conditional is logically true. This explanation then justifies the text's concern with validating arguments (respectively, with proving logical truths) alone. But the standard logic text is concerned only with a case for which the principle of bivalence holds, and we must resist the temptation to extrapolate its teachings to other contexts.

From 3 and 6 we obtain finally the equivalence

7. *A* presupposes *B* if and only if:
   (a) *A* necessitates *B*,
   (b) (*not-A*) necessitates *B*.

Thus we have an explication of *presupposition* in terms of standard semantic notions.

Before turning to the task of making our account precise, we may briefly consider two questions. The first notes that we use 'if . . . then' in, for example, 3 and 6; how is this conditional to be understood? For the sake of clarity, I propose that this be understood as the ma-

[2] This relationship is usually called "semantic entailment," but in the present context that terminology might be confusing. Note that the 'whenever' does not refer to times, but to possible situations.

terial conditional. This yields at once the limiting case that, if B is universally valid, every sentence presuppose it. The second question notes that every case of implication or presupposition is a case of necessitation; does the converse hold? This question is not necessarily decidable at this point, because we have not given a complete account of implication (nor presume to be able to do so). But in section III we present a case of necessitation which is not one of presupposition and which, we shall argue, is also not a case of implication.

II

There are two reasons for constructing a formal paradigm at this point. The first is to show the possibility of having distinct semantic relations of implication, necessitation, and presupposition. That is, we wish to show that the distinctions we have drawn are not merely verbal distinctions, but are capable of expressing nonequivalent concepts. The second reason is an obvious one: the notions of implication and presupposition play an important role in certain philosophical arguments. A case in point is the derivation of paradoxes such as the *Liar,* which we shall examine below. A formal paradigm will give us a means of testing such reasoning.

This may suggest that we mean to construct an axiom system. This is not so. Instead, we intend to serve the two purposes indicated by outlining the construction of a certain kind of artificial language. This kind of language is meant to provide a model for the kind of discourse in which presuppositions are important. Because of our explication of presupposition in terms of necessitation, we shall be able to concentrate on the relations of implication and necessitation alone. Our construction will have to be such that: a case of necessitation need not be a case of implication or of presupposition; a case of presupposition need not be a case of implication; and sentences of which some presupposition is not true, are neither true nor false.

An artificial language has two parts: a *syntax* and a *semantics.* The syntax comprises a *vocabulary* and a *grammar,* which together generate the set of its *sentences.* The semantics defines what I shall call the admissible *valuations*; that is, it delineates the possible ways in which these sentences could be true or false together. (In general, the class of admissible valuations is defined in a rather roundabout way, through a (partial) interpretation of the vocabulary.)

The languages whose construction we shall now outline, we call *presuppositional languages.* By this I mean that in their semantics we make it possible for presuppositions to be made explicit, and we countenance the possibility that, for some of the sentences, presuppositions may fail.

We use 'L' to refer to an arbitrary language of this kind and describe those features which make it a presuppositional language. First, we require that the vocabulary include a negation sign and a disjunction sign (other propositional connectives may be defined in terms of these). I may as well say at once that I will not be satisfied if negation and disjunction do not obey the laws of classical logic. With respect to logic I am conservative: I would resist any imperialism on behalf of classical logic; I would not accept the idea that it is applicable to all contexts or that it is sufficient for all (important) purposes, but on the other hand I have no inclination to change it.

Secondly, the vocabulary may contain a special sign for implication. The material conditional ($A \supset B$) may be defined as ($\sim A \vee B$). But some other sign may be specified as *the* implication sign. Here we shall require that the semantic assertion that A implies B be so construed that, at least, it is true only if ($A \supset B$) is logically true. That is, logical truth of the material conditional is a provisionally acceptable explication of implication, and so is any account under which implication entails this logical truth.

In the semantics, we need two preliminary notions. First, the semantics must specify a relation N of necessitation ("nonclassical necessitation") to cover those cases which cannot be cases of implication. How this relation N is specified in the semantics need not concern us here.[3]

The second notion is that of a *classical valuation.* This is an assignment of truth (T) and falsity (F) to sentences, which disregards the possibility that a sentence might suffer from a failure of presupposition. Given our remarks about *or* and *not,* we know three things about classical valuations:

8. If v is a classical valuation for the language L and A, B are sentences of L, then
   (a) $v(A) = \text{F}$ or $v(A) = \text{T}$,
   (b) $v(\sim A) = \text{F}$ if and only if $v(A) = \text{T}$,
   (c) $v(A \vee B) = \text{F}$ if and only if $v(A) = v(B) = \text{T}$.

Classical propositional logic is clearly *sound* with respect to classical valuations, since classical valuations correspond to rows in the ordinary truth tables. But of course, there may be sentences of L that are assigned T by all classical valuations (*classically valid* sentences of L) which are not theorems of the propositional calculus. (From here on we shall say that a valuation *satisfies* A if it assigns T to A, and that it *satisfies* a set X if it satisfies every member of X.)

[3] In this paper, N will be taken to be a relation of sentences to sentences; the general case of such a relation from sets of sentences to sentences will not concern us here.

In general, no classical valuation will be entirely correct with respect to an actual situation. With respect to an actual situation, the sentences of L are divided into three classes: those which are true, those which are false, and those which are neither. Since a sentence is true if and only if its negation is false, it follows that which sentences are true determines which sentences are false, and hence determines also which sentences are neither true nor false. In other words, in so far as the situation can be described in L, it is determined uniquely by the set of sentences of L which are true with respect to it.[4]

Suppose that G is the set of sentences true in a given situation. What do we know about G? First, no sentence in G has a presupposition that fails. Therefore, the sentences of G can all be true together from the point of view of the classical valuations, and what follows from them from this point of view really does follow. (There are sentences about which the classical valuations are radically wrong, since they cannot accommodate a lack of truth value; but such sentences do not belong to G.) On the other hand, the classical valuations disregard N; hence there are some consequences of true sentences that they overlook. If A belongs to G, and N(A,B), then B is also true; hence also belongs to G. We may sum this up as follows:

9. (a) There is a classical valuation that satisfies G,
   (b) if every classical valuation that satisfies G also satisfies B, then B is in G,
   (c) if N(A,B) and A belongs to G, then so does B.

A set of sentences G for which 9(a) holds we call *classically satisfiable,* and if 9(b) and 9(c) hold for it, we call G a *(necessitation-)saturated* set. So our conclusion is that the set of sentences true in an actual situation is a classically satisfiable, saturated set.

It is at this point that we can discuss the admissible valuations. An admissible valuation is to correspond to a possible situation, such that it assigns T to the sentences true in that situation, F to the sentences false in that situation, and does not assign a truth value to those sentences which are neither true nor false in that situation. On the basis of our present discussion, we can therefore say the following about admissible valuations:

10. An assignment s is an *admissible valuation* for L only if there is a classically satisfiable, saturated set G such that:
   (a) if A is in G, then $s(A) = \mathrm{T}$,
   (b) if the negation of A is in G, then $s(A) = \mathrm{F}$,
   (c) otherwise, $s(A)$ is not defined.

[4] This can be made precise by using the notion of *model,* but the intuitive notions suffice for our present purpose.

We may note that $(\sim \sim A)$ is in G if and only if A is in G; so a sentence has (or lacks) a truth value if and only if its negation does. Also, since G is saturated, if A necessitates B and B is not in G, neither is A. From this it follows by our characterization of presupposition that if A has a presupposition which is not true, then A is neither true nor false.

There is another important point about admissible valuations, which establishes that whatever logic is sound with respect to the classical valuations is also sound with respect to the admissible valuations. This is the point that an admissible valuation represents what is common to a certain set of classical valuations. The assignment s characterized by 10 is identical with the supervaluation induced by the set G in question—which notion is defined by:

11. The *supervaluation induced* by G is the function that
    (a) assigns T to A if all classical valuations that satisfy G assign T to A,
    (b) assigns F to A if all classical valuations that satisfy G assign F to A, and
    (c) is not defined for A otherwise.

And clearly what is common to classical valuations cannot transgress laws that hold for all classical valuations (in particular, the laws of classical propositional logic).

Specifically, the *law of excluded middle* continues to hold:

12. Any sentence of the form $(A \vee \sim A)$ is valid (assigned T by all admissible valuations).

But the *law of bivalence* does not hold: some sentences are neither true nor false.

Reasoning concerning presupposition, implication, or necessitation can be tested through our formal paradigm. With a view to this critical function of the formal paradigm, we shall point out certain features of the notion of presupposition which can easily be demonstrated with its help.

First, the reader will have no difficulty in verifying that there can be distinct cases of necessitation, implication, and presupposition, so that these are three distinct semantic relations. Secondly, we may note that if A presupposes A, then A is never false (it may be true, or neither true nor false). Similarly, if A presupposes $(\sim A)$, then A is never true. If both are the case, if A presupposes a contradiction, then A is always neither true nor false. And if A is presupposed by a valid sentence, then A is valid. (It is clear that the relation N might be such that the language had no admissible valuation: a patho-

logical case which could model only discourse with inconsistent presuppositions.)

III

The first subject to which we shall apply the distinctions drawn above, is *truth*. This is not caprice; the points here made will then play a role in our discussion of the paradoxes of self-reference. The main question before us is: what is the relation between the sentence "It is true that P" and P itself? (We symbolize the former as $T(P)$.) The answer purports to be given by Tarski's principle:

13. $T(P)$ if and only if P.

But what relationship is this "if and only if" meant to indicate? One obvious answer is that it signifies *co-implication*. But this is not necessarily so; for example, one *might* say "If the King of France is bald, then he exists" to signify that the antecedent necessitates the consequent. (This would be a confusing use of 'if . . . then', to be sure, and would most likely indicate a failure to distinguish between implication and necessitation.)

Do P and $T(P)$ imply each other? Principle 13 is used in the "derivation" of the semantic paradoxes, and has also been used to "derive" the law of bivalence. I wish to consider this latter "derivation" here, and shall present it in its shortest form. Given that 13 must be understood to assert a co-implication, we have in particular

14. (a) P implies $T(P)$.
    (b) $\sim P$ implies $T(\sim P)$.

Suppose now that $T(P)$ is not the case; that its denial is the case.[5] Then, by (a) and modus tollens, the denial of P is the case. But this conclusion and (b) lead by modus ponens to: $T(\sim P)$ is the case. So if $T(P)$ is not the case then $T(\sim P)$ is the case. Since our metalinguistic 'if . . . then' is the material conditional, this means that either P is true or $\sim P$ is true: that is, either P is true or P is false. But if the language is a presuppositional language, this conclusion does not hold.

The first, and obvious, reaction to this argument is that the distinction between use and mention has not been observed. Let us try to observe it. Let L be a presuppositional language, and let us form its metalanguage M as follows:

15. (a) sentences of L are first-level sentences of M,[6]
    (b) if P is a sentence of L, then 'P' is a name in M,

[5] This assumption of bivalence for $T(P)$ will be discussed shortly.

(c) 'T' is a predicate of M which is applicable only to names formed by applying quotation marks to sentences of L,
(d) if A and B are sentences of M, so are $\sim A$ and $(A \vee B)$,
(e) A sentence that has any well-formed part beginning with 'T' is a second-level sentence of M,

where '$\vee$' and '$\sim$' are also the disjunction and negation signs of L. We continue to let '$T(P)$' stand for 'T' followed by P in quotation marks. We shall use 'P' and 'Q' to stand for first-level sentences only.

We do not at this point have the full semantics for M, but at least the semantics of the first level must coincide with the semantics of L. But our argument above is easily restated making use of M (and taking as premise only the minimal assertion that P materially implies $T(P)$). But, however we complete the semantics of M, the argument cannot establish that the principle of bivalence holds for first-level sentences, because that is not so.

We have three possibilities before us. We can reject the premises, or we can reject the rules whereby the conclusion is derived, or we can accept the conclusion as formulated by the second-level sentence

16. $$T(P) \vee T(\sim P)$$

of M, but interpret this sentence differently. Let us consider this last alternative first. We remember that a sentence of the form $P \vee \sim P$, which we had been inclined to interpret as saying that P is either true or false, need not be so understood. We could similarly use supervaluations or a many-valued matrix to reinterpret second-level disjunction in such a way that 16 does not say that P is either true or false. The most obvious way to do this is to say that when P is neither true nor false, then $T(P)$ does not have a truth-value either —as opposed to: then $T(P)$ is false. But we shall then have no way of formulating the assertion that a sentence is *not true*. Nor could we add a third level to M in which to formulate this assertion without running into the same problem. That is, assuming that the transition to each higher level obeys the same principle, we would simply get: if P is neither true nor false, then neither is $T(P)$, and neither is $T(T(P))$, and so on. Thus we have here a solution, but it means that M is in some important ways not a model of the metalanguage we have actually been using.[7]

The second possibility, of rejecting the validity of the argument, is not a very pleasant one either. The core of the argument, which can be formulated in M, is

[6] Strictly speaking, the first level of M should be isomorphic to, rather than identical with L; but this makes no essential difference here.

[7] This also answers the question of why we assumed bivalence in the formulation of the argument (see fn. 5).

17. $P \supset T(P)$	
$\sim P \supset T(\sim P)$	contraposition
$\sim T(P) \supset \sim P$	
$\sim T(P) \supset T(\sim P)$	transitivity
$\sim\sim T(P) \vee T(\sim P)$	def. material implication
$T(P) \vee T(\sim P)$	double negation

and is entirely validated by classical propositional logic. Again, I would not be satisfied if this were rejected; such a rejection might throw radical doubt on our own metalogical reasoning. But, instead of rejecting 17, we might reject the original argument on the basis that from "A implies B" we ought not to conclude that A materially implies B, and that the moves in 17 do not apply to implication proper. I would not care to prevent anyone from introducing a new arrow (this has been a popular and, I would claim, harmless pastime), but I do not really think it would be to the point here.

Rather than turn to such radical departures from the standard logical framework, we may consider the possibility that the argument is not sound. As we originally formulated it, the argument proceeded mainly by modus tollens and modus ponens. Only the latter is valid for necessitation in general. Therefore we will have circumvented the problem very simply by interpreting Tarski's principle not as a co-implication but as a *co-necessitation.*

We can now also complete the semantics of M in a very obvious way:

18. (a) If the first-level sentence P is true, then $T(P)$ is true, and otherwise $T(P)$ is false;
    (b) If the second-level sentence A is false, then $\sim A$ is true, and if the second-level sentence A is true, then $\sim A$ is false;
    (c) If $(A \vee B)$ is a second-level sentence, then if either A is true or B is true, then $(A \vee B)$ is true, and otherwise $(A \vee B)$ is false.

Clearly, if A is neither true nor false, then it is a first-level sentence, and so is $\sim A$; if B lacks a truth value as well, then $(A \vee B)$ is a first-level sentence. The principle of bivalence holds for second-level sentences. Classical propositional logic remains sound for the language M.[8]

That this semantics satisfies the principles we have adopted is seen as follows: when P is true, so is $T(P)$, and conversely; hence the two necessitate each other. When P is neither true nor false, then $T(P)$ and $T(\sim P)$ are both false, and $T(P) \vee T(\sim P)$ is also false. When P is neither true nor false, then so is $\sim P$, and $T(P)$ is false;

[8] This is easily seen by embedding the valuations into a three-valued matrix in which the negation of the middle value gets T and the disjunction of two middle values gets F.

hence $(\sim P \vee T(P))$ is false: P does not materially imply $T(P)$. There is, finally, an interesting case of presupposition in M: P does not presuppose $T(P)$, since $\sim P$ does not necessitate $T(P)$. But both P and $\sim P$ necessitate $T(P) \vee T(\sim P)$. Therefore P presupposes its own bivalence; we may call this the ultimate presupposition. Moreover, this is rather a stable result, for to show it we need not appeal to the bivalence of second-level sentences nor to the truth definition for disjunction beyond the unproblematic feature that, when one of the disjuncts is true, so is the disjunction. Not all the results of this section have such stability, as we shall see when we leave the relative security of M.

IV

So far we have assumed that there is a neat division between sentences about sentences and the sentences they are about. I do not mean simply that we have observed the distinction between use and mention. I mean that, in addition, this distinction corresponds to a division of the class of sentences with which we have so far been concerned. That is not the same thing: there is certainly a distinction between loving someone and hating someone; yet it is possible to hate someone you love. So we have not only observed the distinction between use and mention, but assumed that no expression is mentioned in the course of its use.

But this assumption is not a necessary one, and this brings us to the subject of self-reference. The famous Liar paradox is the obvious point of departure. The Liar says "What I now say is false." Clearly this is an English sentence, perfectly grammatical. Yet it can be construed as mentioning itself: the Liar could equally well have said, "This sentence is false" or "The sentence which I now utter is false." The use of this sentence involves mentioning it; hence the distinction between using sentences and mentioning sentences, while a perfectly good distinction, does not correspond to a neat division among sentences.

The paradoxical element appears when we ask whether what the Liar said was true or false. (In the following argument, we assume principle 13 only in the weakened sense of "P and $T(P)$ necessitate each other.") If what he said was true, then what he said was the case; but what he said was that what he said was false. So if what he said was true, then it was false. Similarly we can demonstrate that if what he said was false, then what he said was true. In other words, both the supposition that it was true and the supposition that it was false, lead to absurdity.

This conclusion is itself absurd only on the assumption that what he said must be either true or false. But we are now quite used to

the failure of bivalence; so we simply say: what he said was neither true nor false. The air of paradox is spurious.

Before we begin to feel too smug about this, however, we must face a second paradox, which I shall call the *Strengthened Liar* and which was designed especially for those enlightened philosophers who are not taken in by bivalence. The Strengthened Liar says "What I say is either false or neither true nor false."

If we now ask whether the sentence is true or false or neither, we find that each of these answers is absurd. For example, suppose that what he says is neither true nor false. Then clearly it is either false or neither true nor false. But then what he said was the case. So what he said was true. And now we seem to be properly caught, our sophistication with respect to bivalence notwithstanding.

One move that one might consider, though unhappily, is to say that there is a fourth possibility. This has first of all the unwelcome consequence of facing us with a Strengthened Strengthened Liar (and so on *ad infinitum*). Secondly, it is our desire to conform to classical logic. The principle that has for us replaced bivalence is:

19. $$T(P) \vee T(\sim P) \cdot \vee [\sim T(P) \,\&\, \sim T(\sim P)]$$

which is a second-level sentence and is valid no matter what P is, for it is a tautology, a theorem of propositional logic. We can deny $T(P) \vee T(\sim P)$, bivalence, for that is not a tautology, but we cannot deny 19.

However, I have not led you through the labyrinthine distinctions among implication, necessitation, and presupposition merely for its own sake. But before I show how these distinctions may be mobilized to help us see our way through the Strengthened Liar paradox, I must say something about English sentences and our symbols. We use $\sim P$ to stand for the denial of P, and so when P is an English sentence, say "Tom is tall," we generally read $\sim P$ as "Tom is not tall" or "It is not the case that Tom is tall." But if X is, for example, the Liar sentence "What I now say is false," we see that its denial is not expressed by "What I now say is not false." If the Liar were to utter both, he would have to utter them in succession, so the 'now' would refer to two different times. Hence he would not deny his first statement by making the second, the second referring only to itself. Yet X does have a denial—the Liar's audience, or he himself, may respond "That is not so." Similarly, in deriving the absurdity we argue, for example, that if what he says is not the case, then what he says is false—using 'what he says is not the case' to express the denial of what he says. The exact English words which may express this denial will depend on *who* denies it *when*. This is a clear sign that our sym-

bolism is not nearly adequate to give a complete picture of self-referential language.

But what we can do is to isolate the features of the Liar sentence that play a role in the paradox. In doing so, we may use X to stand for what the Liar says, and $\sim X$ for its denial, however (when, by whom) the denial may be expressed. And then we see that we can take the essential features of X that are appealed to in the derivation of absurdity, to be its relations to the sentence $T(\ulcorner\sim X\urcorner)$ which expresses its falsity. In all other aspects, we shall assume our formal paradigm to govern inferences involving X, for this reduces the problem to the familiar case. And we shall have succeeded if we can then show how the derivation of absurdity is blocked.

First, we might take X to co-imply $T(\ulcorner\sim X\urcorner)$ and also $T(X)$, and $T(\ulcorner\sim X\urcorner)$ to co-imply $\sim X$. This is the first way in which anyone is likely to take the problem, and it leads to absurdity. *This shows* that this way of taking it is incorrect. (In any case, we have already seen that some of these implications do not hold.) Let us now be cautious and take these relations all to be cases of co-necessitation. Then we find that bivalence is needed to derive absurdity; and this dissolves the paradox in the way we indicated informally above.

We could also put this as follows: X necessitates $T(\ulcorner\sim X\urcorner)$, but so does $\sim X$. Hence, by our definition, X presupposes $T(\ulcorner\sim X\urcorner)$. We have seen that $T(\ulcorner\sim X\urcorner)$ cannot be true; therefore, X has a presupposition that is not true. This is why X is neither true nor false. In this way, the distinction between necessitation and presupposition leads to a solution of the paradox.

Thus the Liar paradox seems to fit very nicely into our conceptual scheme. But the fact is that it has one feature that does not fit at all. For X is itself an assertion of falsity: the Liar says "What I now say *is false*." And in the preceding section, when we constructed the metalanguage M, we followed the principle that, although bivalence does not hold generally, it does hold for assertions of truth and falsity. We are here taking the Liar sentence to be a first-level sentence, but there are no restrictions on the form of first-level sentences. In particular, this sentence is an assertion of falsity, and it does not have a truth value.

This I shall call the *basic lesson* of the Liar paradox: even assertions of truth or falsity do not in general satisfy the law of bivalence.

Turning now to the Strengthened Liar paradox, we find a sentence Y which co-necessitates its own falsity-or-truthvaluelessness. That is, Y and $(T(\ulcorner\sim Y\urcorner) \vee \cdot \sim T(Y) \,\&\, \sim T(\ulcorner\sim Y\urcorner))$ necessitate each other. Each of the three possible suppositions—that Y is true, that Y is false, that Y has no truth value—necessitates a contradiction. We have here three

valid arguments with a common conclusion; and this conclusion is a self-contradiction. This conclusion is demonstrated if one of these three valid arguments must have true premises. This is tantamount to:

20. $T(Y)$ is true, or $T(\sim Y)$ is true, or $\sim T(Y) \,\&\, \sim T(\sim Y)$ is true.

This looks like the tautology 19, but it is not the same *unless* we persist in the opinion that assertions of truth are themselves always true or false. And this was the basic lesson of the ordinary Liar paradox: that opinion is mistaken.

To put it most perspicuously, 20 is related to 19 as bivalence to excluded middle. We cannot conform to logic and also deny 19, or excluded middle. But we can deny bivalence, and we can also deny even that sentences that begin with 'T' are bivalent. The Strengthened Liar paradox is averted if we hold that $T(Y)$ and $T(\sim Y)$ are themselves neither true nor false. From this it follows immediately that the sentence $\sim T(Y) \,\&\, \sim T(\sim Y)$ also is neither true nor false.

And we can also give a good reason for holding this. As for every sentence, $T(Y)$ necessitates Y. But, by the tautology 19, $\sim T(Y)$ necessitates $T(\sim Y) \vee \cdot \sim T(Y) \,\&\, \sim T(\sim Y)$. The latter in turn necessitates Y. Hence $T(Y)$ and $\sim T(Y)$ both necessitate Y; in our terminology, $T(Y)$ presupposes Y. As we have seen, Y cannot be true. This is why $T(Y)$ is neither true nor false. (Similarly for $T(\sim Y)$.)

At this point it may be instructive to see how we would extend our formal paradigm to accommodate the deviant sentences X and Y. We must be careful again not to extend it in such a way that classical logic is violated, for then our own reasoning might be drawn in doubt. First we place X and Y among the first-level sentences. They bear no unusual relations to other first-level sentences; hence we need not extend the relation N. But we must now add to the semantics of M a relation N* of nonclassical necessitation, and say that X and $T(\sim X)$ bear N* to each other. Similarly Y and $T(\sim Y) \vee \cdot \sim T(Y) \,\&\, \sim T(\sim Y)$ bear N* to each other. Let us call the thus extended language "M*." The admissible valuations of M become the classical valuations of M*. We define 'saturated set of sentences of M*' as before, except that we use N* instead of N. The admissible valuations for M* are supervaluations generated by saturated sets of sentences, as before. Only the sentences belonging to the entourage of X and Y will be affected by this; the others keep their normal truth value (if any). That classical logic continues to hold can be demonstrated as before.

The notion of "level" has been made much less sharp; the relation N* imposes trans-level semantic relations among the sentences. But

there is still a clear and distinct *syntactic* notion of level. We can also add a third level, in which we can, for example, express the fact that $T(Y)$ is neither true nor false. We would do this by extending M* in just the way that we extended L to produce M. But let us not deceive ourselves: we shall not get to a point where we can say everything relevant, and yet not have any presuppositions that could fail. To be presuppositionless may be a regulative ideal in philosophy, but it is not an achievable end.[9]

In conclusion, I should like to describe briefly a further kind of paradox, and attempt to apply my analysis to it. Epimenides the Cretan is reported to have said that all statements by Cretans are false. Clearly, what he said cannot be true. For what he said was said by a Cretan, and hence he has implicitly asserted its falsity. But we can consistently hold that what he said is false. This just means that *something* said by some Cretan is not false. And this is not as implausible as Epimenides seems to have thought. But, as Church has pointed out, whether or not some statement by a Cretan is not false, is a contingent matter. In particular it entails that the statement of Epimenides which I have just described is not the only one ever made by any Cretan. So the world could have played a neat trick on us: this could have been Epimenides' first and last statement, and all other Cretans could have been entirely dumb. In that case, neither could we have held that what he said was false.

This paradox we shall call the *Weakened Liar*. Epimenides said in effect that all other statements by Cretans are false, and also that his own (this very) statement is false. Let his sentence be Z and let the sentence that expresses that all other statements by Cretans are false be Q. Then Church's point is that the denial of Z necessitates $\sim Q$.[10] But Z necessitates $T(Z)$, which necessitates a contradiction; hence Z also necessitates $\sim Q$. Therefore, $\sim Q$ is a presupposition of Z; and if it fails, Z is neither true nor false. But if $\sim Q$ is true, the case is different: Z and Q are such that $\sim Q$ necessitates $\sim Z$. Therefore if $\sim Q$ is true, then Z is false.

This final example intends to show how our analysis can be applied to members of this family of paradoxes other than the two for which it was developed explicitly. On the other hand, we have not supplied a general theory of self-referential language. This is the familiar lament that we have no (sufficiently general) formal pragmatics. But the paradoxes of self-reference have been a major ob-

[9] That we envisage only finitely many levels here is not a necessary limitation, and not essential to this point.

[10] In the context of the fact that the statement is made by a Cretan; in the present argument, *necessitation* and *presupposition* are relativized to this assumption.

stacle to such a pragmatics, and we shall be satisfied if we have shown that this obstacle at least is not insuperable.

BAS C. VAN FRAASSEN

Yale University

NOTE. The standard discussions of *presupposition* are those of P. F. Strawson in his *Introduction to Logical Theory* (London: Methuen, 1952), and "On Referring," *Mind,* LIX, 235 (July 1950): 320–344. Strawson's account has been critically discussed a.o. by W. Sellars, "Presupposing," *Philosophical Review,* LXIII, 2 (April 1954): 194–215, and G. Nehrlich, "Presupposition and Entailment," *American Philosophical Quarterly,* II, 1 (Jan. 1965): 33–42. The former was answered by Strawson in the same issue of that journal, pp. 216–231, and the latter corrected by Nehrlich himself in "Presupposition and Classical Logical Relations," *Analysis,* XXVII.3, 117 (January 1967): 104–106. The literature on *implication* is now too voluminous to be summarized; we refer only to A. R. Anderson and N. D. Belnap, Jr., "The Pure Calculus of Entailment," *Journal of Symbolic Logic,* XXVII, 1 (March 1962): 19–52. Correspondence with Peter Woodruff, Wayne State University, has helped me to become clearer on the distinction between implication and presupposition.

The basic idea for our treatment of *presupposition* was suggested by Karel Lambert, University of California at Irvine, to whom I am much indebted, in a discussion of section VII of this author's "Singular Terms, Truth-value Gaps, and Free Logic," this JOURNAL, LXIII, 17 (Sept. 15, 1966): 481–495. The language of free logic as there described is a presuppositional language in the sense of sec. II of this paper. The notion of *supervaluation* in this paper is a generalization of that presented in "Singular Terms, Truth-value Gaps, and Free Logic." Supervaluations have since been used in R. Meyer and K. Lambert, "Universally Free Logic and Quantification Theory" (forthcoming), and in B. Skyrms' comments on J. Pollock's "The Truth about Truth" at the APA (Western Division) meetings in Chicago on May 4, 1967 (see below).

The distinction between excluded middle and bivalence, apparently first made by Aristotle, was introduced in this century by the Polish logicians: see S. McCall, "Excluded Middle, Bivalence, and Fatalism," *Inquiry,* IX, 4 (Winter 1966): 384–386. That we are in good company here is witnessed by Quine, who mentions Paul Weiss, Yale University, as having been brought "to the desperate extremity of entertaining Aristotle's fantasy that 'It is true that p or q' is an insufficient condition for 'It is true that p or it is true that q'." (*The Ways of Paradox,* New York: Random House, 1966, p. 21). The nature of validity and the consequence relation in presuppositional languages is explored in my "Presuppositions, Supervaluations, and Free Logic" in the forthcoming Festschrift in honor of Henry Leonard (ed. by Karel Lambert).

Section III is an improvement and extension of sec. VIII of "Singular Terms, Truth-value Gaps, and Free Logic." The three-valued matrix mentioned in sec. III was suggested by footnote 23 of E. Sosa, "Presupposition, the Aristotelian Square, and the Theory of Descriptions" (mimeographed, University of Western Ontario, 1966); the question of the adequacy of such matrices for classical propositional logic is discussed in A. Church, "Non-Normal Truth-Tables for the Propositional Calculus," *Boletin de la Sociedad Matematica Mexicana,* X, 1–2 (1953). I should like to thank John Heintz, University of North Carolina, for helpful suggestions concerning the use of this matrix.

The literature on the paradoxes of self-reference is also huge; this paper has profited most from A. N. Prior, "On a Family of Paradoxes," *Notre Dame Journal of Formal Logic,* II, 1 (January 1961): 16–32. Also suggestive were N. Rescher, "A Note on Self-referential Statements," *ibid.,* V, 3 (July 1964): 218–220; and D. Odegard, "On Weakening Excluded Middle," *Dialogue,* V, 1 (September 1966): 232–236. Bryan Skyrms, University of Illinois at Chicago Circle, has developed independently a solution to the Strengthened Liar paradox through the use of

supervaluations, which he presented in the paper mentioned above. His account is in some ways very similar to ours, but the syntax of the artificial language used is quite different; it may lead to a much finer analysis of the self-referential language used in the paradox than we have provided.

Finally, I should like to acknowledge my debt to stimulating discussions and correspondence on the semantic paradoxes with V. Aldrich, University of North Carolina, and G. Nakhnikian, Wayne State University, and to thank R. Clark, Duke University, and P. F. Strawson, University College, for their encouraging comments on an earlier version of this paper.

PART II

PRAGMATICAL SELF-REFERENCE

PRAGMATIC PARADOXES

By D. J. O'Connor

PHILOSOPHERS have spent a good deal of time and trouble in elucidating the so-called "logical paradoxes". And although their efforts have not yet been completely successful, these paradoxes are now a good deal less puzzling than they were when they were first propounded. But there is another class of paradoxes which has received less attention, partly no doubt because they do not appear at first sight to raise any interesting technical questions of logic or to point the way to new technical developments. Nevertheless, these "pragmatic paradoxes" as they have been called, are worth examination, although I shall not do any more here than draw attention to some of their characteristics and commend them to the attention of philosophers.

Consider the following case. The military commander of a certain camp announces on a Saturday evening that during the following week there will be a "Class A blackout". The date and time of the exercise are not prescribed because a "Class A blackout" is defined in the announcement as an exercise which the participants cannot know is going to take place prior to 6.0 p.m. on the evening in which it occurs. It is easy to see that it follows from the announcement of this definition that the exercise cannot take place at all. It cannot take place on Saturday because if it has not occurred on one of the first six days of the week it must occur on the last. And the fact that the participants can know this violates the condition which defines it. Similarly, because it cannot take place on Saturday, it cannot take place on Friday either, because when Saturday is eliminated Friday is the last available day and is, therefore, invalidated for the same reason as Saturday. And by similar arguments, Trursday, Wednesday, etc., back to Sunday are eliminated in turn, so that the exercise cannot take place at all.

Now though there is an obvious fault of definition in this case, the fault is not a fault of logic in the sense that the definition is formally self-contradictory. It is merely pragmatically self-refuting. The conditions of the action are defined in such a way that their publication entails that the action can never be carried out. Now why should philosophers be interested in this sort of situation? It seems to me that there are a number of examples of such paradoxes which can arise in philosophical discussions and which deserve the attention of philosophers even if the rather frivolous example I have just given does not interest them.

If I say "I do not exist", my statement is L-false as the word "I" in a given context functions as a proper name. But suppose I say "I remember nothing at all". This is not logically self-contradictory. And it is not merely a false factual statement like "X remembers nothing" where X is giving accurate answers to questions about his past life. But it is like an L-false statement in

that it could not conceivably be true in any circumstances, because I must at least remember the proper use of the English phrase " I remember nothing at all " in order to be able to use it significantly. Thus, *prima facie*, it is a peculiar sort of false statement which is neither logically false, nor yet merely factually untrue.

It is not difficult to multiply instances. For example, when I say, on a given occasion, " I am not speaking now " I am uttering a false statement of a totally different character from, say, " Churchill is not speaking now " when I am listening to his broadcast. Yet it is not a statement which is L-false or one which raises by self-reference logical puzzles of the same sort as " I am lying now ".

Here is one further example of a slightly different type. Suppose I say " I believe there are tigers in Mexico but there aren't any there at all ". This statement, being of the form " p and q ", can be true only if both its components are true. And a curious result follows from this. If I say " I believe there are tigers in Mexico " (p) " but there aren't any there " (q), it is possible for p and q both to be true, *only if I am lying* when I utter " q ". For even if there are *no* tigers in Mexico, the fact that I believe the opposite entails that when I utter " q " I *intend* the statement to be false, whether it is in fact false or not.

There is a feature common to the last three paradoxes I have mentioned. They are all statements in the first person which refer to the contemporary behaviour or state of mind of the speaker. In other words, they are all statements involving what Russell calls " egocentric particulars "[1] and Reichenbach calls " token-reflexive " words.[2] That their peculiarities are closely connected with this can be seen from the fact that the peculiarities disappear if we substitute " you " or " he " for " I " or allow the statement to refer to past or future conditions of the speaker. But not all pragmatic paradoxes are of this kind, and it seems to me that it would be worth while for philosophers to pay a little more attention to these puzzles than they have done up to now even if their scrutiny does no more than make a little clearer the ways in which ordinary language can limit and mislead us.

D. J. O'CONNOR.

[1] *An Enquiry into Meaning and Truth*, ch. vii.

[2] *Elements of Symbolic Logic*, ch. vii (" Analysis of Conversational Language ", para. 50).

MR. O'CONNOR'S "PRAGMATIC PARADOXES"

by L. J. Cohen

IN MIND of July, 1948, Mr. D. J. O'Connor drew attention to four statements constituting what he called "pragmatic paradoxes". The peculiarity of these is that apparently, although they are not formally self-contradictory, they cannot conceivably be true in any circumstances: *e.g.*, "I remember nothing at all" (where I must at least remember the proper use of the English sentence "I remember nothing at all"). In connexion with these paradoxes it is worth comparing some other statements with those mentioned by Mr. O'Connor. For instance, if I say "I remember something", apparently this statement cannot conceivably be false (for I must at least remember the proper use of the English sentence "I remember something"). But it does seem that I can intend it as a factual statement about my contemporary state of mind. Thus apparently on the one hand we seem to have statements which are not self-contradictory but cannot conceivably be true, and, on the other, statements which are not analytic but cannot conceivably be false. I wish to suggest that these paradoxes arise out of an ambiguity (not infrequently recognised) in the word "statement".

In one of the two senses of "statement" which I propose to distinguish it is an event-expression, like "motion", "laughter", or "physical training". Of anything which is called a "statement" in this sense it is legitimate to ask "When and where did (does, will) it happen?". In another sense, however, "statement" is a logical expression, like "entailment", "non-contradiction", or "type fallacy". And of anything which is called a "statement" in this sense it is absurd to ask "When and where did (does, will) it happen?". We can avoid the ambiguity by using "utterance" in the former sense, instead of "statement", and "proposition" in the latter. Thus utterances will, by definition, be events in space and time. Some utterances occur in speech, for instance, others in writing or in silent thought. And propositions will, by definition, not be events in space and time. I do not mean that propositions are subsistent entities, for I do not know what these are. But "utterance" and "proposition" stand to each other in such a relationship that it makes sense to ask, *e.g.*, "When should I utter the proposition 'The cat is on the mat'?", or "What proposition was his utterance intended to communicate?".

Now there is nothing to debar propositions from describing utterances just as they describe other events. "The cat is on the mat", "All his utterances are in English", and "None of my utterances take place between 2 a.m. and 8 a.m.", can all be regarded as propositions. Accordingly there is nothing in principle to debar propositions from being such that they can be verified or falsified by their own utterance.

If I now utter the proposition " I remember nothing at all ", I should indeed be uttering a false proposition. But if I have my utterance recorded for a gramophone and the record is played at my funeral, the proposition uttered might then be true. Thus the proposition " I remember nothing at all " can conceivably be true. And the proposition " I remember something " can conceivably be false, if uttered in similar circumstances. But this can only be recognised if we distinguish " proposition " from " utterance " in such a way that the same proposition may be uttered in different circumstances at different times.

Similarly, the proposition " I am not speaking now " would be false if I spoke it aloud. But it would be true if I thought it silently to myself. And Mr. O'Connor himself mentions circumstances in which the proposition " I believe there are tigers in Mexico but there aren't any there at all " would be true : it would be true if I am lying when I utter " but there aren't any there at all ".

Mr. O'Connor notes that three of his four paradoxes resemble each other in being " statements in the first person which refer to the contemporary behaviour or state of mind of the speaker ". But he also mentions another paradox which does not contain any such egocentric particulars. " The military commander of a certain camp announces on a Saturday evening that during the following week there will be a ' Class A blackout '. The date and time of the exercise are not prescribed because a ' Class A blackout ' is defined in the announcement as an exercise which the participants cannot know is going to take place prior to 6 p.m. on the evening in which it occurs. It is easy to see that it follows from the announcement of this definition that the exercise cannot take place at all. . . . The conditions of the action are defined in such a way that their publication entails that the action can never be carried out." I suggest that, although this paradox differs from his others in not involving egocentric particulars, it resembles them in arising from a proposition that can be falsified by its own utterance. The proposition in this case is " A ' Class A blackout ' will take place during the following week ". This proposition is rendered false by its public announcement : it might be true if the camp commander told nobody of his intention to hold the exercise.[1]

[1] If the camp commander intended to stage a surprise exercise on one day during the week and yet wanted to warn his troops of his intention, he would have to make an announcement somewhat like one or other of the following : Either " One day next week there will be a surprise exercise. A surprise exercise is an exercise about which, unless it takes place on the last day of the period for which you are warned, you will be in doubt as to when it is to happen until 6 p.m. on the evening in which it occurs." Or " One day next week there will be an exercise. Unless it takes place on Saturday you will be in doubt as to when it is to happen until 6 p.m. on the evening in which it occurs." In the former case he utters a prediction and a definition, in the latter two predictions. Owing

It would be interesting to know if there are any pragmatic paradoxes—statements which are apparently not self-contradictory but not conceivably true, or factual but not conceivably false—that do not arise from failing to distinguish between the "utterance" and "proposition" senses of "statement" (or similar words) when the proposition in question can be verified or falsified by its own utterance.

L. JONATHAN COHEN.

to the irreversibility of the time series, if it is known that an event will take place on either $t1$ or $t2$ or $t3$ or . . . tn, it will only occur as a surprise (in the ordinary sense of "surprise") if it happens on either $t1$ or $t2$ or $t3$ or . . . $tn - 1$. For this point I am indebted to a discussion with E. A. Gellner and K. Rankin.

PRAGMATIC PARADOXES

by Peter Alexander

IN MIND of January this year, Mr. Cohen attempts to deal with the "Pragmatic Paradoxes" put forward by Professor O'Connor in July, 1948. It seems to me that both Mr. Cohen and Professor O'Connor have overlooked the real confusion in one of these paradoxes, at least.

Professor O'Connor gave as an example the statement "I remember nothing at all" which is not logically self-contradictory and not merely factually false; and although it could not conceivably be true because I must at least remember the proper use of the English sentence "I remember nothing at all" it is not like the statement "I am lying now", which raises puzzles by self-reference.

Mr. Cohen considers that the paradox alleged to be involved in the statement "I remember nothing at all" arises from an ambiguity of the word "statement". Thus, within the meaning of this word he distinguishes "utterance" from "proposition" in such a way that the same proposition can be uttered at different times, so that the proposition "I remember nothing at all" might conceivably be true if a record of it made by me were played at my funeral, because then I might truly be said not to remember the correct use of the English sentence "I remember nothing at all", or indeed anything else. (It should be noted that this post-mortem performance is necessary to Mr. Cohen's elucidation and that no other would serve.)

I shall not here dispute the necessity of making this distinction between "proposition" and "utterance" for some purposes, but I do want to question whether it is of any use in connexion with the statement "I remember nothing at all".

It seems to me that the egocentric particular "I" can only be used sensibly by a living person of himself.[1] When the gramophone record is played at my funeral whom or what does "I" denote? Normally "I" denotes the person speaking. Here the gramophone "utters" the proposition and yet "I" could not sensibly be said to denote the gramophone; at the most it could denote the person who made the record *at the time at which the record was made* and certainly not my corpse at the time at which the record was played. Possibly if "I" were regarded as a proper name irrespective of context or if the gramophone could point to my corpse there might be some reason for supposing it to denote my corpse but even here there are difficulties. It would be necessary for the "I" denoted to be the same as the "I" denoted by me when I made the record; and part of what "I" denotes when I use it of my living self may be held to be the very thing whose absence is now demanded—my remembering or my "consciousness" or at least

[1] Except in quotation, *e.g.* "He said, 'I remember nothing at all'"—a usage which is irrelevant here.

my ability to use or understand the English language. It is probably not sensible to say of my dead body either that it remembers something or that it remembers nothing: it has ceased to be the kind of thing that, as far as we can know by any normal means, could remember—or be denoted by "I". On the other hand, as Mr. Cohen points out, if the gramophone were to utter the proposition "He (or that) remembers nothing at all", even if it could be taken to be a sensible statement about my corpse, this would embody no paradox whether I were dead or alive.

Mr. Cohen's solution then appears to be unsatisfactory in two ways for (1) we should not think it sensible to use "I" to denote a corpse because the use of "I" is restricted to use by, and of, living persons, and (2) even if we could sensibly use "I" of a corpse the position is simply reversed, but not cleared up, for in this situation the proposition "I remember something" is not self-contradictory but could not be verified (although it might conceivably be true) and the proposition "I remember nothing at all" is not analytic but could not be falsified (although it might conceivably be false).

It seems to me that the "paradox" contained in the statement "I remember nothing at all" arises not from an ambiguity of the word "statement" but from an ambiguity, quite as familiar, of the word "remember".

We can distinguish at least two senses of the word "remember". I can remember (1) how to ride a bicycle or how to swim and (2) riding a bicycle through Brighton in 1939 or swimming in the sea at Bangor in 1949—and it is obvious that I use "remember" in two quite distinct senses. In the first sense I retain an acquired skill or habit, and it is in this sense that I remember how to construct English sentences; in the second sense I recall, by means of images or otherwise, events from my past experience. Through this distinction lies the way out of the paradox which is, *prima facie*, involved in the statement "I remember nothing at all". We can recast the statement to allow for this distinction so that it might read—

"I am still able to construct an English sentence correctly but I cannot recall my past experiences".

Now the proposition expressed by the subordinate clause of this sentence can be verified or falsified and it is not falsified by its utterance. The proposition expressed by the main clause *is* verified by its utterance or by its entertainment—and the puzzle is one of self-reference like that of "I am lying now".

On the other hand, the proposition "I am no longer able to construct an English sentence correctly" would be falsified by its utterance but is not comparable to the paradox thought to be embodied in "I remember nothing at all" by Professor O'Connor and Mr. Cohen. The paradox contained in the proposition "I am no longer able to construct an English sentence correctly" arises, like that contained in "I am lying now", from self-reference. The proposition is falsified by the structure of the sentence which expresses it.

I think that this analysis of the proposition "I remember nothing at all" is consistent with what would normally be meant by anyone who uttered it.

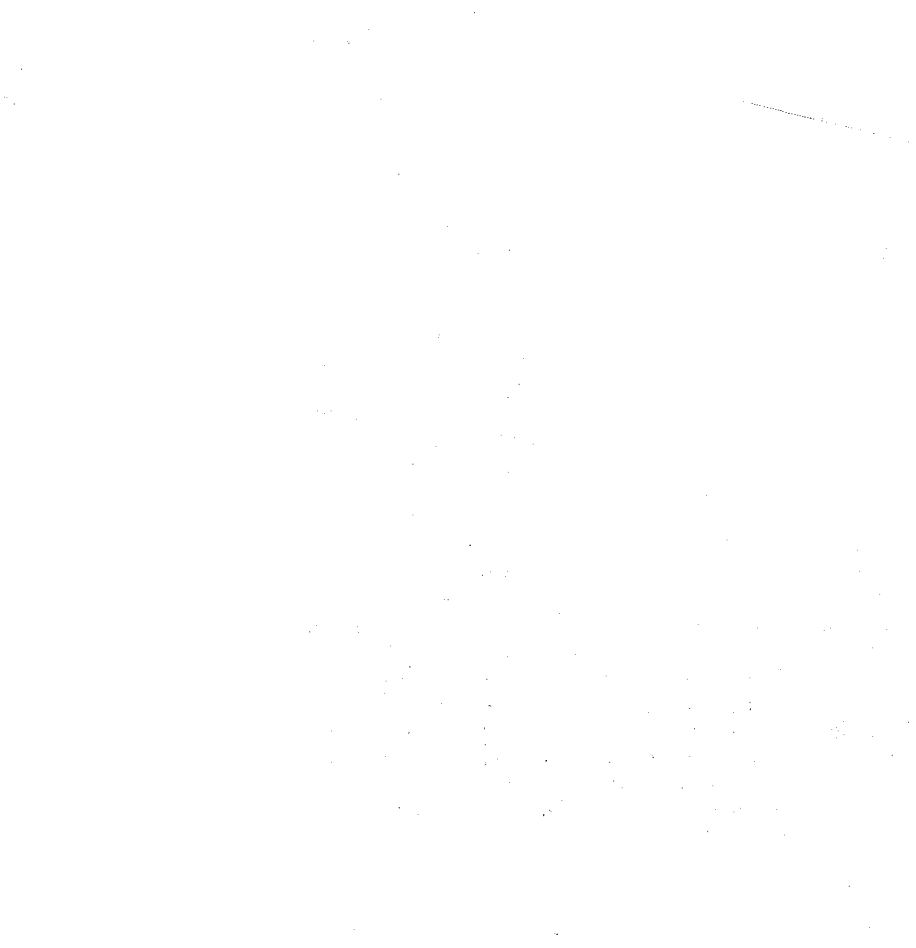
Now this "pragmatic paradox" differs from the one involved in the statement "I am not speaking now" because it does not seem possible to remove this paradox by distinguishing two senses of "speaking". Perhaps Mr. Cohen's distinction is satisfactory here for the proposition "I am not speaking now" need not be false if merely entertained (unless we subscribe to the view that "entertaining" is equivalent to "sub-vocal speaking") so that the uttering of it falsifies it and, as Mr Cohen points out "there is nothing in principle to debar propositions from being such that they can be verified or falsified by their own utterance". Thus this proposition differs from "I am lying now" and the alleged paradox in "I remember nothing at all" for these would be falsified by their mere entertainment. If "entertaining" *is* taken as equivalent to "sub-vocal speaking" or "sub-vocal uttering" then the proposition "I am not speaking now" is peculiar in just the way that Professor O'Connor suggested—but we have no need to accept this view of "entertaining".

Now both paradoxes with which I have dealt differ from Professor O'Connor's "Class A Blackout" paradox, although he seems to think they have much in common. If they are different then the "Class A Blackout" paradox, though interesting, it is of no great concern, unless there are others like it which might arise in normal philosophical discussion. But it seems to me that any announcement of an intention is implicitly recognised to be conditional on the possibility of carrying out that intention. Even if I make a simple statement like "I will go to the cinema to-morrow" I mean, although I do not state, that I shall do so if I am not in any way prevented. Thus Professor O'Connor's statement, which can be abbreviated to read "A 'Class A Blackout' will be carried out next week" *ought* for completeness, to read "If the conditions of a 'Class A Blackout' can be realised, a 'Class A Blackout' will be carried out next week." Now this seems to raise no other difficulties than are raised by any conditional statement whose condition is unrealisable, like, for instance, "If I can live without air I will not breathe all day to-morrow". Of course, I might be able to live without air to-morrow but, similarly, men might cease next week to be able to realise that if the blackout had not occurred by Friday it must occur on Saturday, and then the condition would be realisable.[1] Any problems raised by these statements do not appear to be similar to those raised by the other statements with which I have dealt nor to be properly called "paradoxical".

PETER ALEXANDER.

Leeds University.

[1] Professor O'Connor stresses the point that no logical contradiction is involved. If that is so, his statement is not like the statement "To-morrow things equal to the same thing will not be equal to one another". There is no logical contradiction in men ceasing to be able to *see* a relation.



FUGITIVE PROPOSITIONS

By AUSTIN DUNCAN-JONES

MANY of the propositions we commonly entertain have an odd yet obvious feature which does not seem to have been much remarked on. This feature belongs to many empirical propositions which imply the existence of some event or class of events. It does not belong to *a priori* propositions, or to genuine universals of law. And it is that the proposition in question cannot be entertained twice. In other words, the usual assumption that every proposition has what D. R. Cousin calls " spatio-temporal neutrality "[1] is mistaken.

This feature of propositions is made clear at once by formu-

[1] *Proc. Aris. Soc.*, 1948–49, p. 153, " Propositions ".

[Continued on next page]

lating them in a tenseless language, in which date is expressed by a variable or constant time coefficient. Instead of

" Brutus killed Caesar ", (1)

we might say

" Brutus kills Caesar at T ", (2)

in which the present tense is used in a timeless sense. But (2) is not an adequate rendering of (1) unless we explain the meaning of " T ". The obvious way to do so is as follows.

"For some t, Brutus kills Caesar at t, and t is before now", (3)

where "now" is the proper name of a moment. Whenever I use (3) on different occasions to express a proposition which I entertain, one element of the proposition entertained is different : therefore on each occasion the proposition entertained is different.

This is not a self-contained oddity. For it follows that no proposition susceptible of the kind of analysis illustrated in (3) can ever be a matter for deliberation or controversy. If historian A uses (3) on a given occasion, and historian B wishes to disagree with him, B will not be able to express the proposition he rejects, for the moment he needs to name will have passed.

To produce this absurdity I have to some extent distorted the customary use of the word " proposition ". For given that the names are used in the same way it is part of the customary usage that people who say " Brutus killed Caesar " on different occasions *do* express the same proposition. But it is also part of the customary usage that if, when a sentence is used on two occasions, some part of it has different meanings on the two occasions, the sentence cannot express the same propostion on each occasion. And if S1 and S2 are sentences with the same meaning, and if S1 expresses different propositions on different occasions, then S2 expresses different propositions on different occasions. Therefore, if (3) is an adequate rendering of (1), the accepted usage of " proposition " is not consistent.

If we seek to retain the convention that historians *do* always express the same proposition by (1), we shall have to allow sentences with different meanings to express the same proposition. In that case a counterpart of our paradox will arise about meanings. Let us assume that (3) is an adequate rendering of (1), in the sense that anyone who uses (1) could have used (3), on the same occasion, to express what he meant by (1). Then if A uses (1) on a given occasion it will never be open to B on a later occasion to reject what A meant by (1).

I suppose we could remove the difficulty by a heroic semantic device which I shall not develop in detail. We could define

B's rejection at t_2 of A's assertion at t_1 in a special way. We could say that B does not assert the contradictory of A's proposition, but asserts a proposition of higher order, to the effect that a certain proposition, which he describes as having been asserted by A at t_1, and as having a certain content, was false. Such an unwieldy piece of machinery seems disproportionate.

It might be argued that propositions mentioning a more or less definite date can be freed from paradox by non-hierarchical methods. For example, in

"Caesar died in 44 B.C.", (4)

the reference to the birth of Christ can be interpreted as an abbreviation for a circumstantial history told in the Gospels. And we might take it as part of what is asserted in (4) that one and only one event, throughout the whole of past and future time, corresponds to this history. The whole proposition can then be taken as asserting a time interval between two described events, without reference to the moment of assertion. I do not think this analysis is plausible. For it implies that (4) might refer indifferently to past or future events. It seems to me that from what we mean by (4) it is deducible, that Caesar is dead.

I think in recent philosophy analysis of the meaning of time expressions has been either neglected or dealt with perfunctorily, although a generation or two has passed since, in other ways, philosophers began in Alexander's words, to "take time seriously". We lack a simple standard method of expressing time relations more explicitly than they are expressed in (1), which will enable us to keep as much as possible of the customary usage of the phrases "same proposition" and "different proposition". I have carried my account of the customary usage only far enough to display my difficulty, and have ignored much of its subtlety.

The time with which we are concerned is the crude time of common sense.

University of Birmingham.

PRAGMATIC PARADOXES AND FUGITIVE PROPOSITIONS

by D. J. O'Connor

BOTH Mr. L. J. Cohen (MIND, January, 1950) and Mr. Peter Alexander (MIND, October, 1950) have clarified some important points connected with the puzzles which I have mentioned in my discussion note on pragmatic paradoxes (MIND, July, 1948). But there are still some points on which I would like to disagree with them.

(*a*) I shall consider, first of all, the paradox involved in the statement " I remember nothing at all ". (This paradox was suggested to me in discussion by Professor F. B. Ebersole.) Mr. Alexander points out that this sentence can be interpreted in two different ways : (A) " I cannot recall my past experiences " ; (B) " I can no longer construct an English sentence correctly ". He adds that (B) is the only interpretation which involves a paradox and that the paradox arises from self-reference like that involved in " I am lying ". In support of this opinion he says that " the proposition is falsified by the structure of the sentence which expresses it ". Now it seems to me that this is a mistake. " I am lying " or, to take the more convenient form in which Grelling discusses it, " this sentence is false " is certainly contradictory through self-reference and is, as Ramsey showed, a *semantic* paradox involving and arising from the naming relation. But Ebersole's paradox, in the form in which Mr. Alexander discusses it, does not involve self-reference in the same sense at all. The sentence (B) does indeed *exhibit* by its grammatical and syntactical structure the fact that it is false but it does not *refer* to itself. There is a very important difference between sentences which are falsified by their own structure and sentences which predicate falsity of themselves. This difference may be shown in two ways.

(i) Propositions of the second type involve semantic paradoxes which can be made manifest by a formal proof that, where p is the proposition in question, p is equivalent to not-p. Such proofs have been given by Grelling (MIND, October, 1936) and Ushenko (*Problems of Logic*, London, 1941, p. 60). But it is not possible, so far as I can see, to prove formally that p is equivalent to not-p, where p is Mr. Alexander's proposition (B).

(ii) We can remove Ebersole's paradox by expressing the same proposition in different words as : " Me no more savee talk English good " or " Je ne sais plus construire correctement une phrase en anglais ". But we cannot remove Grelling's paradox by translation. The contradition in " this sentence is false " remains when we translate to " cette phrase-ci est fausse ".

The reason for this is that Grelling's paradox is a contradiction arising from the relation between the index-sign " this " and what is denoted by the sign whereas Ebersole's is pragmatic in that it arises from the relations between signs and their *users*. I think

that Mr. Alexander may have been misled by a secondary feature of this paradox, namely, that the sentence which embodies the paradox refers to the natural language in which the sentence is expressed. (The statement " I am not now using the English language " embodies a similar contradiction.) Such sentences are sign-vehicles or sentence-tokens which refer to themselves but do not, like the semantic paradoxes refer to the proposition which the sentence-token expresses. That is why the contradiction is dissipated by translation, since translation changes the sign-vehicle or medium of expression without changing the proposition or the meaning expressed.

It seems to me obvious that the main difficulty involved in the use of such statements as " I remember nothing at all ", " I am not speaking now " or " I never speak English " arises from the use of the token-reflexive word " I " together with the present tense of the verb. For the paradox vanishes if we substitute " he " or " you " for the pronoun or change the tense of the verb.

(*b*) I think that this point will help in discussing the solution proposed by Mr. Cohen. He points to the distinction between two senses of the word " statement ", an utterance or sentence-token and a proposition. There is nothing to prevent a proposition from referring to an utterance and " accordingly there is nothing in principle to debar propositions from being such that they can be verified or falsified by their own utterance ". He concludes that the paradox involved in such statements as " I am not speaking now " can be seen to result from the relations between a proposition and its utterance. The utterance may be used to falsify the proposition, and in this case we have a pragmatic paradox. But it may also be used in such a context that it does not falsify the proposition. For example, I may think silently " I am not speaking now " ; or I may make a record of myself saying " I remember nothing at all " which is afterwards played as part of my funeral service when the sentence may very well be true.

But here again there is a difficulty which arises from the use of token-reflexive expressions which I do not think that Mr. Cohen has taken fully into account. It seems to me that the distinction between proposition and utterance cannot be usefully applied to statements which refer to the contemporary behaviour or experience of the speaker. Token-reflexive expressions function in use as variables whose values are specified by the context in which the utterance occurs. Thus its meaning, like the validity of a bus ticket, is exhausted by the single occasion of its use. Such statements are, as it were, incomplete. But normal *complete* statements like " All swans are white " or " Nero died in A.D. 69 " resemble season tickets in retaining their validity through a series of successive presentations. We might state the comparison more exactly by saying that the normal relation between a proposition and the utterances which express it as a one-many relation, whereas in the case of

a statement referring to the contemporary behaviour or experience of the user, the relation between proposition and utterance is one-one. This is an analytic consequence of the fact that our experience is temporal.

Thus Mr. Cohen's explanation does not entirely dispose of the problem because in the case under discussion the type token distinction cannot be usefully invoked. Mr. Duncan Jones has recently discussed some analogous problems which arise from the use of what he calls " fugitive propositions ". It is an " odd yet obvious feature " of such propositions " that the proposition in question cannot be entertained twice " (*Analysis*, October, 1949). The class of Mr. Duncan Jones' fugitive propositions contains as a sub-class what I have called above " incomplete statements ". Dismissing the " Class A Blackout " paradox, which I think Mr. Alexander has exorcised successfully, we could sum up by saying that all pragmatic paradoxes arise from the use of a sub-class of fugitive propositions. Such paradoxical propositions belong therefore the the genus " fugitive propositions ". What is their *differentia?* First, that they shall refer to the contemporary experience or behaviour of the user of the proposition. Secondly, as Mr. Cohen has suggested, on the occasion of the use of such an incomplete statement, the utterance must falsify the proposition. Thirdly, the proposition must be falsified by attributing to the user contemporary behaviour or states of mind which are incompatible with those entailed by uttering or otherwise stating the proposition in question. But I am not sure that this is the whole story.

D. J. O'CONNOR.

University of Natal.

PRAGMATIC IMPLICATION[1]

C. K. GRANT

Introduction.

THE PURPOSE of this paper is to clarify some of the logical problems raised by certain uses of the word "imply" which, although very familiar in ordinary language, have not been adequately investigated by philosophers. There have been numerous references to this type of implication in recent philosophical writings. Some of these are listed below.[2] However, there does not exist, to my knowledge, any account of this concept in its own right; this deficiency I hope to remedy, in part, in the following remarks.

We may begin with a brief explanation of the context in which the problem arises; its importance will be made plain, I hope, in the course of the discussion. Since Aristotle, logicians have investigated the rules which determine the types of inference that can be validly made from one proposition to another. Recently it has become clear that propositions are theorists' abstractions. For example, they are so defined that they must be either true or false; but this distinction cannot be applied to many of the sentences which we utter, such as questions, commands, etc. Hence a logical investigation into language as it is in fact used must take account of the actual utterances that are made, that is, linguistic *practice*. This has a complex logic of its own, which differs from that of formal deductive systems.

It is now generally recognized that the complexities of the concept of implication cannot be completely represented in the terms either of material implication or of strict implication or entailment. It has also been noticed that the senses of "imply" which cannot be treated along these lines are those in which the word is applied to overt utterances or statements, rather than to propositions or propositional functions. This concept of implication is concerned not with the formal relations of propositions but with the properties of sentences in use: nonetheless, as I shall try to show, in this use of the word we

[1] An early version of this paper was read to Professor R. Carnap's Faculty Seminar on logic in the University of Chicago in 1951. The present article is substantially the same as a paper read to the Northern Universities Philosophical Society Conference in 1956.

[2] P. F. Strawson: "On Referring," *Mind*, 1950. D. J. O'Connor: "Pragmatic Paradoxes," *Mind*, 1948. Y. Bar-Hillel: "Analysis of 'Correct' Language," *Mind*, 1946; "Indexical Expressions," *Mind*, 1954. P. T. Geach: "Russell's Theory of Descriptions," *Analysis*, 1950. N. Malcolm: "The Verification Argument," in *Philosophical Analysis*, ed. Max Black. Max Black: "Saying and Disbelieving," *Analysis*, 1952. Reprinted in *Problems of Analysis*. P. H. Nowell-Smith: *Ethics*, pp. 80–4.

refer to relations between statements that are governed by rules which are logical in character, although they are different from the rules of formal logic.

Our investigation of the senses in which it is proper to say that statements have implications will involve some discussion of the complementary but neglected question: what are the rules governing inferences of this type? In other words: what is it to make an inference from a statement rather than from a proposition?

Two preliminary observations must be made about the terminology that I shall use.

(i) I have adapted the name "pragmatic implication" from the earlier of Mr. Bar-Hillel's two papers already cited, where he calls it "pragmatical implication." This expression is, in my view, to be preferred to "contextual implication," which is the term employed by Professor Nowell-Smith in his book *Ethics*. Nowell-Smith's expression fails to bring out the fact that this sense of "imply" is concerned with linguistic *practice;* this is done by "pragmatic implication," which is a useful parallel expression to "pragmatic paradox."

(ii) The meaning that I shall give to the words "speaker" and "hearer" requires some explanation. By the former I mean the utterer of any word or sentence, whether written or spoken, or the user of any non-verbal symbol or combination of symbols that could be translated into a word or sentence. The word "hearer" is used very widely to refer either to the person to whom a statement is directly addressed, as in a conversation or letter, or to the class of people who understand a statement not directly addressed to them, e.g. anyone who overhears a statement directly addressed to another, the audience of a speech or broadcast, the readers of a book, etc.

Non-Symbolic Pragmatic Implication

I shall first consider the cases in which actions other than the employment of words or other symbols, may be said to imply propositions. It is in this sense that we can say that the act of selling something implies the proposition that the seller owns the article or is empowered by the owner to sell. Nearly always, of course, the acts of offering for sale and of selling will involve the making of statements. I am treating these as non-verbal actions for the following two reasons:

(i) Although the use of language is customary in performing acts of this kind, it is not essential; buying and selling can quite well take place between people who do not understand each other's language. It is doubtless essential that each must have his own language in which to calculate price, etc.

(ii) When statements are made in the course of performing these actions, they are relatively unimportant elements in the act as a

whole. To sell something *may* involve saying things like "What will you give me for this?" but it is clear that to offer something for sale is to do far more than to use these words. Thus I mean by a "non-symbolic act" any act in which the use of symbols is not an essential part, and any act which is not definable as a use of certain symbols.

Two types of non-symbolic pragmatic implication must be distinguished.

I. The first may be illustrated by the actions of selling and of marrying. Either of these actions implies a number of propositions, both affirmative and negative. The act of selling something not only implies the proposition noted above but also, as the law has it, that "the goods are free from any charges or incumbrances in favour of a third party not known or declared to the buyer at the time the contract is made."[1] Similarly, the act of marrying implies the propositions that the agent is not already married, is over the age of sixteen, etc.

What is the relation between these actions and the propositions which they imply? This may be outlined as follows: The propositions implied by these actions are those propositions that are *entailed* by any proposition that correctly describes the actions as acts of selling, marrying, etc. This is to say, if the propositions pragmatically implied by these actions were not true, then the acts would not be acts of selling or marrying. They would be described as "invalid," or "null and void." The American folk hero who sold Brooklyn Bridge to an immigrant did not in fact sell the bridge at all; a bigamous marriage is not a sort of marriage, it is a "form" of marriage. The beliefs of the agent are quite irrelevant to the question of whether or not a given act pragmatically implies a certain proposition. The agent may sincerely believe the propositions that are pragmatically implied by his action, e.g. that he is the owner of the goods which he is offering for sale; but if he should be mistaken in these beliefs, then his act is not an act of selling. In this sense of "imply," then, there is no practical absurdity in performing an action whose pragmatic implications are false; it is a logical impossibility. The relation may therefore be defined in the following way: An act A pragmatically implies proposition p when the proposition "Somebody is performing or has performed A" entails p.

All entailments presuppose a system of definitions; entailments of this kind, and the pragmatic implications connected with them in the way outlined, are based upon a system of law and convention. These differ in time and place, and so do the corresponding pragmatic implications of the actions to which the laws and customs apply. These latter may be regarded as either specifications of the necessary conditions for the performance of these actions, or as definitions of the

[1] Cheshire and Fifoot: *The Law of Contract*, p. 97.

actions themselves. To marry in Saudi Arabia is not to imply that one is not already married, and to marry in Kentucky is not to imply that one is over sixteen years of age. No such great variations exist in the laws of buying and selling.

There are some cases where it is unclear whether this notion of implication is applicable or not. If two people, A and B, become engaged, does this pragmatically imply that neither of them is already married to someone else? That is, is it logically impossible for A and B to become engaged if either or both is married? This is unanswerable on account of the vagueness of the notion of "engagement." This has not been legally defined, and hence it is not clear whether the mere intention to marry at some time in the future does or does not count as an engagement.

II. The second type of non-symbolic pragmatic implication cannot be defined in terms of entailment; here the proposition that is pragmatically implied by an action is not a logical consequence of the proposition that an act of a certain kind has been, or is being, performed. Two preliminary points must be made here.

(i) We do not, in ordinary language, use the word "imply" in the sense that I shall now discuss. This is not of much importance, for as I shall try to show we do infer propositions from actions, and furthermore these inferences are genuine pragmatic inferences—that is, they are neither logical deductions nor ordinary empirical inferences.

(ii) The following remarks apply only to intentional, as distinct from involuntary actions; and I shall be concerned only with those intentional actions that are designed to bring about a certain state of affairs, rather than those actions that are performed "for their own sake."[1] These need not be qualitatively different sorts of actions; singing "O Sole Mio" as a serenade illustrates the former, while singing it in the bath is an example of the latter.

Generally speaking, any action of the sort that I am now considering will pragmatically imply: (*a*) psychological propositions about the "state of mind" of the agent, i.e. his beliefs, expectations, and wishes; (*b*) empirical propositions about what may be called the "natural outcome" of the action, in other words its most probable consequences. This is what we call the "point" of the action; (*c*) propositions about the empirical conditions necessary for the achievement of the natural outcome of the action. Suppose we observe a man planting seeds in the soil. From this we can infer the psychological

[1] It is not always easy to draw this distinction. Hiccuping, sneezing, blushing, etc., are all *generally* involuntary and hence cannot be purposive; but some people are said to be able to blush at will. The criteria by means of which we decide whether any given blush is intentional are doubtless complex; however, the relevant point here is that once the distinction is made, we can ask of an intentional blush "What was the point of doing that?" a question that cannot be asked about an involuntary blush.

propositions "He believes that the seeds will grow"; "He hopes that the seeds will grow"; "He believes that the conditions are suitable to bring about the natural consequence of the action, viz. the germination of the seeds." More particularly, we can infer that he believes there is adequate irrigation, drainage, etc. We can also infer the non-psychological empirical propositions "The seeds will grow" and "The conditions are suitable for the seeds to grow." We may call the former the *psychological* pragmatic implications and the latter the *propositional* pragmatic implications of the action.

We are entitled to draw these inferences only because we have formed an opinion about the natural consequences of the action; and this is possible only if we know the relevant empirical laws connecting the action with its most probable consequences. This, however, does not render the pragmatic inference itself a matter of induction; we do not judge that Smith intends the germination of the seeds, or that the seeds will grow, on the ground that most people that we have observed to plant seeds have intended them to germinate, or that most seeds that have been planted do germinate. I shall say more about this later.

This example of planting seeds is liable to be in some respects misleading, for it is pretty clearly an action which is, in general, performed for only one purpose. If it should fail in this we can conclude with confidence, though not always correctly, that the act was pointless and perhaps even absurd. Thus the absurdity or irrationality of an action may be defined in terms of the falsity of the propositions which it implies in this sense of "imply." The majority of human actions are more complex than this example of planting seeds because they can be located in a large number of different purposive schemes. Hence we cannot always be sure of the point of the action, and *a fortiori* cannot conclude that it must be pointless or absurd if it does not lead, and is not intended to lead, to a certain consequence. Only by examining the action in its context can we rule out various purposes which the action, considered in itself, may be designed to achieve.

I have said that although pragmatic inferences of this kind rest upon factual knowledge of human purposes and of the connections between the means that are adopted and the ends that are sought, to make such an inference is to do something different from merely applying an ordinary inductive generalization to a particular instance. It is rather, I suggest, to make a tacit appeal to the concept of rational action. The principle that people usually intend the most probable consequences of their actions is not empirically or inductively grounded, although of course it is both empirically confirmable and falsifiable. The principle is, I think, part of the definition of a rational action; and it is a necessary condition of understanding

human action that it should be rational in this sense. We cannot hope to understand the actions of others, in an important sense of "understand," unless we presuppose that on the whole agents intend the probable consequences of their actions and furthermore believe that empirical conditions will permit the occurrence of these results. This principle is the ground of psychological pragmatic inferences. A somewhat different but complementary conception of rational action underlies propositional pragmatic inferences: roughly, if we regard others as rational agents then we *ipso facto* assume that they have acquainted themselves with the facts of the situation in which they are acting; this entitles us to infer not only the proposition that specifies the probable consequence of the action, but also propositions concerning the conditions for bringing about this result. It is relevant here to recall that the law infers *mens rea* from an action by reference to the criterion of what a reasonable man intends, and could be expected to know, in his situation.

I have argued that the propositions pragmatically implied, in this sense, by an action, may be regarded as the conditions of acting rationally in that particular situation. If these propositions are false, the action is pointless. Now one can imagine many circumstances in which a man may plant seeds believing that they will germinate but hoping they will not—e.g. a disgruntled farm employee; in this case the psychological pragmatic inference will be wrong, but the propositional pragmatic inferences may well be true; or he may plant the seeds believing and hoping that they will not germinate—perhaps he is a botanist, or has made a bet that no seeds will grow there. In that case both types of pragmatic inference will be mistaken.

But note that these are exceptional cases; they deviate from the normal. In ordinary circumstances the falsity of the pragmatic implications of an action renders it pointless and perhaps absurd. Thus if a man is not a botanist, dissaffected labourer, gambler, etc. etc., then his planting of the seeds is pointless and absurd unless he believes that the seeds will germinate. It is also, though to a lesser degree, absurd if the seeds fail to germinate, but I think we would use such strong language as "absurd" only if the agent had no evidence that the seeds would germinate. E.g. the groundnut scheme. It may be argued that here I am merely defining the pragmatic implications of an act in such a way as to entail its pointlessness if the propositions pragmatically implied by it are false. This is exactly what I am doing, viz. explicating the concept of rational action by reference to the pragmatic implications of actions. Hence it is not surprising that a tautology results; but it is, I think, an important and interesting one.

It is because reference is made to this notion in pragmatic inferences that they cannot be placed in the category of ordinary inductive inferences. Admittedly, purposive action, by definition, involves the

making of means–ends judgments, and this requires on the part of the agent knowledge of the appropriate empirical uniformities; but when the spectator makes a pragmatic inference from such an action he does not appeal to, for example, the generalization "It has been observed that in most cases when people plant seeds they expect them to germinate and they do germinate" as, when he explains a blink, he might appeal to the generalization "It has been observed that when the majority of people sneeze they close their eyes." The essential point here is that once an action is classified as purposive it is thereby related to its end; we cannot then take it out of relation to this end as we do if we describe the act as evidence either for the state of mind of the agent or for the propositional pragmatic implications of the act. The idea of evidence is relevant only after the question "Is this action purposive?" (i.e. purposive in the sense explained), has been asked and answered in the negative.

Symbolic Pragmatic Implication

From one point of view a piece of linguistic behaviour is simply a kind of action, so that in this respect any utterance can be said to have pragmatic implications of the second type discussed in the preceding section. The action of uttering a certain sort of sentence has both psychological and propositional pragmatic implications. In this sense we can say that a man who utters an indicative sentence implies that he believes it and that he wants others to believe it (psychological pragmatic implications); his utterance also has the propositional pragmatic implication that another person is present. Someone who utters an interrogative sentence implies that he wants an answer; that he believes that his interlocutor can give it; and again that another person is present. Someone who utters an imperative implies that he wants obedience, expects to get it, and that there is another person present to carry it out.

These implications are quite independent of both the content and the context of the utterance; they are inferences that we draw from the utterance of certain types of sentence, and depend upon the fact that it is pointless for a speaker to utter these sorts of words unless he has a certain "propositional attitude" to what he is saying, certain expectations of his audience, and an audience. If the pragmatic implications of these verbal performances are false, the speaker can escape the charge of absurdity and irrationality only if he has been using this language in a special context such as reporting a conversation, telling a story, acting, testing a microphone, rehearsing a speech, performing an exercise in elocution, etc. etc.[1]

[1] As Prof. J. L. Austin points out: "We don't talk with people (descriptively) except in the faith that they are trying to convey information." *Other Minds.* Reprinted in *Logic and Language* (Second Series) p. 129.

From this two points emerge.

(i) This consideration applies also to the asking of questions, the issuing of imperatives, etc.

(ii) That it should be carried on in this way can be regarded as a *definition* of "talking descriptively," or interrogatively, imperatively and so forth. This is an important consideration to which I shall later return.

I have contended that inferences of this type are neither empirical inductions nor, in the ordinary sense, entailments recognizable *a priori*, but rather that they make a tacit appeal, or in law an overt appeal, to the concept of a rational man. This is the notion of an ideal, and to that extent a pragmatic inference is an *a priori* affair; but empirical considerations concerning the context of the action or utterance enter in when we wish to infer in detail what a rational man would have known or intended in that situation. For this reason it may sometimes look as if we are conducting an ordinary empirical investigation; but this is a mistake. The courts of law also conduct detailed enquiries into matters of fact, but the administration of law remains altogether different from the task of finding the facts—the decision of the court is not a factual record. The *a priori*—empirical dichotomy which bedevils philosophy even more than the analytic—synthetic distinction with which it is closely associated cannot accommodate procedures of this sort. We have no right to assume that all exercises of reason *must* take the form of discovering and recording facts or *a priori* inference; it is a gross over-simplification to suppose that thinking must consist either of fact collecting or of logical manipulation. A pragmatic inference is not an *inductive* inference about the psychological states of the speaker or the hearer; nor is it a *deductive* inference, because we are dealing here not with propositions but with utterances.

Two types of implication

In the course of this discussion various qualifications and elaborations will be introduced on the way, so that many remarks should be regarded as provisional and subject to later modification. In this section I shall attempt to disentangle pragmatic inference from another type of inference, and pragmatic implication from another sort of implication. I shall confine myself to verbal behaviour, although as I have pointed out, these considerations will apply to communication carried out by the employment of any symbols.

It is generally supposed that to recognize a pragmatic implication is to make an inference from an utterance to propositions of the following two kinds; (*a*) to the proposition expressed by the sentence which the speaker utters. That is, we infer p from the proposition

"Smith utters sentence S" where S expresses p. (*b*) to psychological propositions of the form "The speaker believes p" where p is the proposition which he asserts.[1] We shall later see fatal objections to this account of pragmatic implication, but at the moment our problem is to distinguish pragmatic inferences from ones more familiar to philosophers, in particular psychological inferences about the state of mind of the speaker. We can and do make genuine pragmatic inferences concerning the "propositional attitudes" of speakers, though we shall see that these are not inferences to the propositions actually asserted but to those presupposed by the assertion. However, even on the false view that we can infer a speaker's belief in p from his assertion of p, it is clear that such an inference could not be simply the application of a psychological inductive generalization, for if it should turn out that the speaker did not in fact believe p we would not for that reason conclude that our pragmatic inference had been wrong. The reason is that an inference of this type rests only on the conception of what a rational man would have believed in making that utterance, so that our error would have consisted either in supposing that the speaker was behaving rationally, or in failing to note that, in the circumstances of the utterance, we were wrong in taking the words literally—perhaps the speaker was reporting a conversation, telling a story, etc.

These psychological pragmatic inferences are often described by saying that an utterance is *evidence* of the beliefs of the speaker.[2] This can be seriously misleading, for many inferences that are made from verbal behaviour are not pragmatic, but causal, inferences. Causal implication may be roughly defined as follows: p causally implies q when each proposition refers to events that are causally connected.[3] In such a case we can say that p is evidence for q; but this is quite diffferent from a pragmatic connection between an utterance and its presuppositions, and so to apply the word "evidence" to both is to blur an important distinction. If human beings were so constituted that whenever and only when they remembered something they blushed, then the proposition "Smith is blushing" would imply "He is remembering"; this would be a contingent causal implication, not a pragmatic one. It is therefore confusing for Strawson to say that the utterance of the sentence

[1] "When I say 'S is P' I imply at least that I believe it." J. L. Austin, loc. cit., p. 143.

[2] For example Bar-Hillel says " 'if X is human and X says S, then X believes S' is (generally) highly confirmed."

"Analysis of 'Correct' Language," *Mind*, 1946 p. 337. This is like saying that in general people do not vote "Aye" unless they oppose the opponents of the resolution—as if occasionally a few people did! More of this later.

[3] Cf. Burks and Copi: "Lewis Carroll's Barber Shop Paradox," *Mind*, 1950, p. 220, and also: Burks: "The Logic of Causal Propositions," *Mind*, 1951.

"The King of France is wise" is evidence that the speaker believes that there is a King of France.[1] This suggests that we have here an instance of an inductive generalization to the effect that most people who utter sentences of this sort also believe that the individual in question exists—which is absurd. If "evidence" is interpreted very widely as meaning "a good reason," then Strawson's remark is quite correct, but only because "good reason" is a portmanteau phrase covering both causal and pragmatic implications. Thus if a court of law wished to determine whether at a certain date Smith believed that there was a King of France, it would be relevant to show that at that time he said "The King of France is wise." Another illustration, which also shows that ought implies can in the pragmatic sense of "imply," is as follows: if it were necessary to establish whether at a given time A believed that B was able to perform a certain act X, it would be relevant to know that at that time A had said to B "You ought to do X." We shall later see additional objections to Strawson's other view that the utterance of the sentence "It is raining" is evidence that the speaker believes it is raining.

The ambiguities of "evidence" and "implication" are dangerous because they mask the fact that a pragmatic inference is not a matter of inductive psychology. A man's utterances are not related to his propositional attitudes in the sort of way that smoke is related to fire. I do not wish to suggest that Strawson believes that they are, but only that his choice of language can lead to that idea. Nor am I contending that it is wrong to say that a man's reason for saying something may be that he believes it to be true; I wish only to deny that in saying this we are referring to an empirical *de facto* connection between belief and utterance such that the one is a good inductive ground for inferring the other. An inductive inference is made when we judge that a man is excited because he shouts a particular sentence; this is not, as Bar-Hillel maintains,[2] a pragmatic inference, because the connection between shouting and excitement is purely contingent. But if we had reason to believe that the man shouted not on account of excitement or pain, etc., but *for a purpose*, then it is open to us to make a pragmatic inference concerning what I have called the "point" of such an action; thus we might conclude that he was shouting to someone deaf, or far away, and so on.

Uttering and Stating

Many difficulties, some of them to do with the notion of evidence, are raised by the famous passage from Moore which constitutes the first philosophical recognition of pragmatic implication. I shall consider Moore's remarks in some detail. In discussing the absurdity of

[1] "On Referring", *Mind*, 1950, p. 330.

[2] "Analysis of 'Correct' Language," *Mind*, 1946, p. 335

statements like "I believe he has gone out but he has not," Moore says: "This, though absurd, is not self-contradictory; for it may well be true. But it is absurd because, by saying 'He has not gone out' we *imply* that we do *not* believe that he has gone out, though we neither assert this nor does it follow from anything that we do assert. That we *imply* it means only, I think, something that results from the fact that people, in general, do not make a positive assertion unless they do not believe that the opposite is true: people in general would not assert positively 'he has gone out' if they believed that he had not gone out. And it results from this general truth, that a hearer who hears me say 'He has not gone out' will, in general, assume that I don't believe that he has gone out, although I have neither asserted that I don't, nor does it follow from what I have asserted, that I don't. Since people will, in general, assume this, I may be said to *imply* it by saying 'he has not gone out,' since the effect of my saying so will, in general, be to make people believe it, and since I know quite well that my saying it will have this effect."[1]

Moore's explanation of the absurdity of this statement is not satisfactory. Its absurdity derives, he says, from the fact that people "generally" believe what they say; if so, then an exception to this generalization would not be absurd but only surprising, In what, then, does the absurdity consist?

In order to answer this we must examine the differences between the verbs "say" and "utter" on the one hand, and "assert" and "state" on the other. These distinctions are paralleled by the differences between the nouns "utterance," "statement" and "assertion." Now Moore uses the words "say" and "assert" as if they were synonymous or near enough in meaning not to matter. Strawson[2] and Bar-Hillel also hold this view. But it is mistaken. Asserting or stating is not just uttering or saying, although of course in order to make a statement we have to say or utter something. There are many circumstances in which I can utter the sentence "He has gone out" without *asserting* that he has gone out; I may be making a joke or translation, quoting someone else, reciting a poem, playing a guessing game, testing a microphone or performing an exercise in elocution. In none of these cases can I be described as asserting, telling, informing, stating or apprising. This is because these utterances occur in contexts (not merely verbal contexts) in which it would be pointless for the speaker to intend his words to be taken literally. This is not a psychological definition of "assertion"; i.e. I am not saying that an assertion

[1] G. E. Moore: *Russell's Theory of Descriptions. The Philosophy of Bertrand Russell*, Living Philosophers Library, pp. 303 ff., Moore's italics.

[2] Cf. "On Referring", pp. 325–6. But note that on p. 330 Strawson betrays some uneasiness about this; he there talks of a man who "seriously utters" a sentence.

is an utterance accompanied by a wish on the part of the speaker to be taken literally. Only sometimes would we regard the announcement "But I didn't intend you to take me literally" as an acceptable disclaimer for having made a statement, even if we believe it. This would not be an adequate defence in a slander case, for example.

Roughly speaking, I am defining "statement" as an utterance made in the sort of context in which it would normally be taken literally; and it would normally be understood in that way because of the absence of those relatively unusual circumstances which permit the non-literal or "free-wheeling" use of language yet to have a point. The criteria to which we appeal in deciding whether particular utterances are statements or not are complicated and have an "open texture," so that borderline cases may be encountered in which the criteria conflict, or in which it may be unclear whether these criteria are applicable or not. I suspect that here, as elsewhere, we may learn a good deal of logic from the law.

In my usage, the word "statement" is not confined to those utterances that are sentence tokens of indicative sentences. It is necessary also to distinguish between the utterance of an interrogative sentence and the asking of a question, and the utterance of an imperative sentence and the issuing of a command. In English the word "statement" is applied only to indicative sentences, but this is a deficiency in the language; accordingly, I shall use the word "statement" to apply to statements in the ordinary sense and also to questions and demands as distinguished from the utterances of indicative, interrogative and imperative sentences. When it is necessary to emphasize the difference between an indicative statement and an interrogative or imperative statement—i.e. when I wish to emphasize the descriptive or true-or-false character of the former—I shall call it an assertion. We shall see that both interrogative and imperative statements have pragmatic implications.[1] If I am correctly described as asserting, stating, etc., then I am also being described as saying what I believe or know, *or else as saying what I pretend to believe or know*. If I am making an assertion and not lying, then it is a tautology that I am saying what I believe or know. Hence the idea that we somehow have to infer (by appeal to some generalization like "People on the whole believe what they say"), from "Smith asserts p" to "Smith believes p" is quite wrong. We would not call it asserting if we doubted whether he believed or knew p, or else was pretending to do so.

But suppose that we do have grounds for this; all that we are given

[1] It is worth observing that so-called "pragmatic paradoxes" can arise in imperative as well as indicative statements. Two examples must suffice.(i) It is alleged that Mr. Arthur Christiansen, the Editor of the *Daily Express*, once sent a directive to his reporters which began "Clichés must be avoided like the plague." (ii) The imperative: "Prepositions are not for ending sentences with." These paradoxes raise interesting questions, but they are not relevant here.

is that Smith utters sentence S, and that S expresses p. Is there not then a problem about the inference from the proposition "Smith utters S" to the proposition "Smith believes p?" No, this is not another problem distinct from the question of whether this utterance of S is a statement or not; once that has been answered, and to do so requires no inferring but an examination of the utterance in its context, there is no more to find out. I do not mean by this the recognition that the utterance is an instance of an induction concerning the correspondence between what speakers believe and what they say, but rather the application to the utterance of those criteria by means of which we determine it to be a statement.

Notice that the question whether or not someone is telling a lie is one that can be asked only if we have already asked and answered affirmatively, the question "Is he making a statement?" Once we have done this, it may be important to determine whether someone is being honest or dishonest with us; then we *do* have to make inferences from his averted eye, shuffling feet, past record and so forth—but by inference we cannot establish whether he is purporting to tell the truth. If this is correct, it is wrong to argue, as does Prof. Black,[1] though somewhat inconsistently with the rest of his theory, that a speaker's assertive tone of voice can be *evidence*, i.e. the basis of an inference, regarding his beliefs. This is like saying that we "infer" that a man is a clergyman from his dog-collar; this is not *evidence* that he is a clergyman, it is a conventional sign that he is a clergyman—he is representing himself as a clergyman. We do not, although we could, discover by experience that wearers of dog-collars are clergymen; we do not observe a correlation between dog-collars and Holy Orders any more than we establish empirically the fact that most people believe what they assert. If we did, it would be in order to say that the assertion, or the tone of voice, etc., are *evidence* which entitles us to make the inference to belief.

This may be brought out in another way. Consider the expression "I confidently assert . . ." This is obviously a redundancy. No one would say "I tentatively, hesitantly or diffidently assert . . . ," and so there is no room for the phrase "a confident assertion." If someone is diffident, then he is not asserting but guessing, surmising, suggesting, and so on. It is part of the meaning of "assert" (and of "assertion"), and of "state" (and "statement"), that people believe, are confident in, what they assert and state, or pretend to be so.

There would be a different absurdity in saying "I guess he has gone out but he hasn't," for if I can assert that he hasn't gone out then I am not in a position to guess that he has. We can bet on certainties but we cannot guess for or against them. This impossibility is not psychological but logical, for "guessing" *means* taking a chance.

[1] *Problems of Analysis*, pp. 43–4.

These considerations apply not only to assertions, that is, a class of sentence tokens of indicative sentences, but also to all statements in the wide sense that I have given the word; if a man is described as asking a question or issuing a command, as distinct from uttering an interrogative or imperative sentence, he is *eo ipso* being described as seeking an answer or obedience.

There are, I think, two reasons why it has been supposed that there can be an inference from someone's assertion of p to his belief in p. (*a*) The assumption that "utter" and "assert" mean roughly the same.[1] (*b*) A recognition of the fact that we frequently raise questions like "Why did he say that?" or even "What made him say that?" and that if for some reason we cannot ask the speaker outright, we often have to make some kind of inference. But, of course, only in very exceptional circumstances would we accept "Because he believed it" as an answer to this question. If he was asserting and not lying we know perfectly well that he believed what he said; what the question seeks is either the grounds for this belief, or the speaker's motive in expressing it. These are both distinct from what I have called the "point" of an action, in this case the making of a statement. Thus if A informs B that B's wife is unfaithful to him, we may ask why A said that, wanting to find out what grounds A had for believing what he said, or what moved him to tell B—perhaps he wished to wound him, or to persuade him to leave an awful woman. In neither case is the question concerned with what I mean by the point of the statement; this is simply the most probable consequence of the act of making the statement, viz. that B will believe something about his wife.

We can now begin to see more clearly the reasons for the absurdity of the statement "I believe he has gone out, but he has not." This should be compared with saying "Placet—but I'm not voting." The expression "placet" records a vote in essentially the same kind of way that the indicative phrase "he has not gone out" records a belief. There is a temptation to say that pragmatic absurdities of this kind are merely very obvious lies that could never take anyone in; but this temptation should be resisted. Not only because a form of words of this sort is not a statement and hence cannot be a lie, but because it confuses pragmatic absurdities with something different. An obvious and futile lie is told if a seated man says, in an "assertive tone of voice," "I'm not sitting down," or sings and immediately says "I'm not singing." The absurdity here lies in the relation between the statement and the non-linguistic context to which the statement refers, whereas a genuine pragmatic absurdity is evident from the

[1] This error was criticized, though on somewhat inadequate grounds, in *Some Notes on Assertion*, by C. Lewy; reprinted in *Philosophy and Analysis* (ed. Macdonald), p. 120 ff.

words of the utterance alone, as Mr. Hare has pointed out in his review of Nowell-Smith's *Ethics*.[1]

I have suggested that Moore's diagnosis of the absurdity of statements of the form "*p*, but I don't believe *p*" is mistaken because an overt infringement of the empirical generalization "Most people believe what they say (or assert)" could give rise only to surprise on the part of the hearer; we cannot in this way account for the utter bewilderment occasioned by a pragmatic absurdity.

We must distinguish between two dimensions in which the utterance of such a sentence is absurd. These correspond to the utterance considered simply as an action on the one hand, and as a misuse of language on the other.

(i) Various kinds of actions, whether they involve the use of symbols or not, may be absurd and pointless.[2] Suppose a man does two things which for empirical reasons are mutually inhibiting; the performance of one prevents the natural consequence of the other, and vice versa. Thus a man who fans a fire with bellows while simultaneously pouring water on it is doing something absurd and pointless in this sense. In some respects this is analogous to saying "*p*, but I don't believe *p*," because the natural consequence of asserting *p* is not only to induce hearers to believe *p*, but also to suppose that the speaker believes it—a consequence which is inhibited by the rider "but I don't believe *p*." A speaker cannot expect anyone to take both halves of the sentence equally seriously.[3] Notice that the absurdity arises only if the assertion of *p* is part of the same statement as the assertion of disbelief in *p*. I might say or signal very slowly "He has gone out" and while doing so notice his hat and coat, and then add "but I don't believe he has." There is nothing queer about this, I am merely correcting an earlier statement by a later one. A pragmatic absurdity is committed by a single conjunctive assertion and not a sequence of assertions.

(ii) This idea that a pragmatic absurdity consists in performing two mutually inhibiting actions is only part of the explanation. This is clear from the fact that we would not describe the utterance of

[1] PHILOSOPHY, 1956.

[2] The view that the absurdity of this sort of utterance arises from its pointlessness was first advanced by Mr. A. M. MacIver. Concerning utterances of the form "*p*, but I think not-*p*," MacIver says "the second half of what I say makes the *saying* of the first half *pointless*" (MacIver's italics). *Some Questions about "Know" and "Think," Analysis* 1938. Reprinted in *Philosophy and Analysis*, ed. M. Macdonald, p. 95.

[3] Black raises the question: "Can we even imagine what it would be like for the utterance in good faith of "*p*, but I don't believe *p*" to have a point? op. cit., p. 49–50. The whole problem lies in the analysis of the phrase "uttering in good faith." It is worth noting that "*p*, but I don't believe *p*" *could be* interpreted, in certain circumstances, as a description of a change in the speaker's psychological state, viz. the transition from belief in *p* to doubting it.

such a sentence as a very obvious lie, nor would we admit that it expresses, or could express, an extremely stupid mistake about a matter of fact—rather, we would say that it could arise only as a result of the speaker's ignorance of the language. I shall return to this point later; I mention it now only to show that a pragmatic absurdity is a kind of *linguistic* aberration. What exactly is it that is odd about this sort of language?

The answer to this is, I think, simple. If, as I have argued, "assert" means "believe or pretends to believe," it follows that anyone who asserts p (i.e. anyone who would be correctly described as asserting p) cannot at the same time assert that he does not believe p, i.e. cannot be described as *asserting* disbelief. This is a logical impossibility which derives from the relations between "assert," "believe" and "pretend to believe."[1] This is confirmed by the fact that although it is possible to pretend to believe, it is logically impossible to pretend to assert.

We can now see that Strawson's account of pragmatic implication is deficient in two respects. His view is that a statement S, which consists in the assertion of p, pragmatically implies q when q must be true before p can be either true or false.[2]

(i) The definition is too narrow. In the same way that an assertion is a declaration of belief, so is a question normally a request for information, and a command a request for obedience. A speaker pragmatically implies, in the proper sense, those propositions whose falsity would render the acts of asserting, questioning or commanding pointless. We may say either that *he* is implying these propositions or that *what he says* implies them. The fact that questions and imperatives have pragmatic implications (or "presuppositions") shows that we cannot define the concepts of pragmatic implication or presupposition in terms of either the truth of, or the speaker's belief in, what he says, for questions and commands cannot be true or false, and hence it makes no sense to talk of believing or disbelieving them.

[1] If I am right to take "assertion" as the fundamental notion in resolving this paradox, we can see the deficiency in Prof. Black's interesting treatment of the problem. He distinguishes between an explicit assertion and its "signification"; a sentence in a certain mood "signifies," but does not state, the speaker's belief (*Problems of Analysis*, p. 52 ff). In "p, but I don't believe p" the assertion of disbelief contradicts the signification of belief. Such an utterance is a misuse of a conventional sign of belief; but it has been shown that we do not learn rules for the use of sentences over and above the rules of particular words and stock phrases. (R. Willis: "Prof. Black on 'Saying and Disbelieving'," *Analysis*, 1953.) Black introduces this unsatisfactory notion of a rule of signification because he states the problem by considering a speaker who *says* p; this is too vague—is he just uttering a sentence or making an assertion? If this question is asked, rules of signification become redundant.

[2] Cf. "On Referring", and also A. J. Baker: "Presupposition and Types of Clause," *Mind*, July 1956, p. 371.

(ii) When restricted to assertions, Strawson's definition will not do, for a reason other than that mentioned earlier. We have seen that Strawson, Austin and Moore all maintain that the assertion of *p* implies that the speaker believes *p*. We may call the proposition "The speaker believes *p*," *q*.

It is not difficult to show that it is impossible to reconcile this with the Strawsonian view that statement S, which asserts *p*, presupposes *q* if *q* must be true before *p* can be either true or false. For if *q* is "the speaker believes *p*," then "S pragmatically implies *q*" entails that the speaker must believe *p* before his assertion of it is to be either true or false. This is absurd, because the applicability of the true-false distinction to a statement does not depend upon the speaker's belief, or disbelief, in it. It *would* be correct to say "X asserts *p*" entails "X believes *p* or is pretending to," but this is an *a priori* truth deriving from the logical grammar of "assert" and cognate verbs such as "state."

There are other kinds of sentence whose utterance, although pragmatically peculiar, is less absurd than the utterance of "*p*, but I don't believe *p*." One of these is the sentence which puzzled Bradley: "The soul is not a fire-shovel." It is always pointless to make this statement. If it is made to someone who can understand it, this is a redundant thing to do, because anyone who can understand it *ipso facto* knows that it is true, in a different sense from that in which understanding a tautology is also to see that it is true. A hearer who had even an inkling of the meanings of "soul" and "fire-shovel" would not need to be told that the soul is not a fire-shovel. If, on the other hand, the hearer did not have an inkling of the meanings of these words, then it would be pointless to make the statement to him on this count. However, with sufficient ingenuity one can devise some sort of employment for the most bizarre sentence. Prof. D. Emmet has put the following case to me. After reading the Bible someone may reflect as follows: "I see that I must love my enemies; I love with my soul: and I am told to heap coals of fire on those who hate me—but this is impossible since my soul is not a fire shovel." Here the sentence occurs in a soliloquy and not a statement addressed to another, so that I am not sure that my argument is substantially affected.

There have been, I thing, three main obstacles in the way of those philosophers who have discussed pragmatic implication. The first is the common belief that this is a "very special and odd sense of 'imply.' "[1] There is a serious mistake here; in ordinary language "imply" means "pragmatic implication." It is the logical theorists in their talk of strict and material implication who give the word special and odd senses. The O.E.D. gives 1581 as the date of the earliest use of "imply" as meaning "pragmatic implication": "he

[1] P. F. Strawson: "On Referring", p. 330.

that forebydethe a thynge to be done in after tyme, doth he not covertly emplye that the same was done before?" The other two factors, I suggest, have been a preoccupation with pragmatic absurdities of the kind discussed by Moore, together with confusion about the correct definition of "assertion."

If it is a *statement* that is under consideration, there can (logically) be no doubt that the speaker believes what he says, or is pretending to believe it—and although it may not always be a simple matter to decide whether a given expression of belief is a fake, there are well known and fairly clear criteria to which we appeal in trying to do so. If we should have reason to believe that the utterance in question is not a statement, that is, should not be taken literally in the context, then there can be no question of the utterance possessing pragmatic implications.

If this argument has been on the right lines, we have established the negative point that a speaker who asserts or utters an indicative sentence is not implying that he believes it, a speaker who states or utters an interrogative or imperative sentence is not implying that he wants an answer or obedience. What, then, is a pragmatic implication?

The answer to this is as follows: A statement pragmatically implies those propositions whose falsity would render the making of the statement absurd, that is, pointless. This follows analytically from the definition of "statement," for part of the meaning of "statement" as distinct from "utterance" is that the former, in its context, is understood literally, and hence implies those propositions which, if false, would make it pointless to utter the sentence as a statement. I have tried to show that a statement is an utterance which is either intended to be, or would normally be, taken literally; therefore to describe a given utterance as a statement is also to describe the speaker as representing himself as a rational man. Such a man:

(*a*) Intends the most probable consequences of his actions.

(*b*) Believes or wishes, etc., what he is stating, for normally no purpose is served by representing oneself as believing or wishing what one does not believe or wish. In the utterances of a rational man, there is a correlation between his "propositional attitudes" and his linguistic performances.

(*c*) His beliefs in those propositions pragmatically implied by his statements are well-founded, i.e. more likely to be true than not. As we have seen, this is the foundation of propositional pragmatic inferences.

Therefore what is pragmatically implied by the assertion of p cannot be p itself, nor can it be "The speaker believes p"; what is implied is another proposition or propositions, q and r. The assertion

"Smith is returning tomorrow" pragmatically implies "Smith has gone away" because there would be no point in making the assertion unless Smith had gone away. The same proposition is implied in the same way by the imperative "Find out where Smith has gone" and the interrogative "Where has Smith gone?" It is also implied, of course, that there is such a person as Smith.

The general formula for this type of implication is: the statement S pragmatically implies proposition q. If S is an assertion, it will be the assertion of some proposition other than (not equivalent to) q. It may be objected that to make a statement is merely to perform an act, and that an act cannot stand in a logical relationship with a proposition or anything else, Accordingly we may reformulate the pragmatic relation as follows: it holds between *propositions* of the form "So and So makes statement S" and a class of other propositions which must be true if the making of the statement is not to be absurd. I think there is nothing wrong with this account, but it certainly does not conform to customary usage, for ordinarily we would say that So and So, in the very act of making the statement, is implying propositions q and r. A man who asks the question "Where do Smith's children go to school?" is in so doing implying the propositions "Smith has children" and "Smith's children go to school." It now seems that it is not propositions that have pragmatic implications, but rather those human actions that consist in the statement—making use of sentences, to use concise but potentially very misleading language. I see nothing wrong with this account either, and it seems quite consistent with also saying that pragmatic implication holds between propositions of the form "So and So makes statement S" and a number of other propositions that are pragmatically implied, though not entailed, by this proposition. The assertion "There is a seal in the bedroom" pragmatically implies "There is a bedroom"; "The bedroom is bigger than the seal"; "There is at least one seal in the house"; and this because unless the speaker were lying, there would be no point in his making the statement unless the implied propositions were true.

That a statement pragmatically implies those propositions whose falsity would render the making of the statement pointless is, as we have seen, a logical consequence of the definition of "statement." It is important, however, to see that the particular propositions enumerated above are not *entailed* by the proposition "Smith asserted that there was a seal in the bedroom" because, of course, he may have been lying or mistaken, etc.

One consequence of this is that a pragmatic inference is quite different from an empirical investigation into the beliefs of the speaker. This may be shown in another way. The propositions "Smith's assertion of p pragmatically implies q" and "Smith's

assertion of p pragmatically implies that he believes q" are both compatible with the proposition "Smith does not believe q." It is not uncommon for people, through ignorance of the language, carelessness and so forth, to make statements of whose pragmatic implications they are unaware. Many social *gaffes* are of this nature. Hence we cannot define "pragmatic meaning" in terms of the intentions of the speaker, as is attempted by C. H. Langford in "The Notion of Analysis in Moore's Philosophy." Cf. *The Philosophy of G. E. Moore* (Living Philosophers Library), p. 332.

There are types of pragmatic implication which rest upon verbal conventions; these may concern either the form of words or the verbal stress of the utterance. The following examples illustrate these two cases.

(*a*) The assertion of p together with a definite descriptive phrase or proper name may pragmatically imply p in respect of the remainder of the class excluded by the descriptive phrase, or individuals other than those denoted by the proper name. There is a church in Nottingham whose notice board announces "There is no colour bar in this church"; the pragmatic implication of this has been recognized with distaste by other religious bodies. An imperative statement may have a pragmatic implication of this kind. An example of this, also by coincidence theological in topic, was recently reported in the Bournemouth local paper, which recounts that a young lady outside church X met the congregation as they left and said to them "Go to church Y, they are all Christians there."

(*b*) Although Cook Wilson pointed out long ago that verbal stress can be of logical importance, this is a notion that has not been investigated. Stress may have pragmatic implications. Thus if I say "He *said* that he was detained at the office," I may properly be said to imply that he was not detained at the office, and that he lied when he said that he was, because I am reporting his statement by means of a vocal inflexion which is conventionally used to express disbelief. This must not be taken to mean that there is here anything like a "rule of signification." The purpose of stressing that a man *says* something is to distinguish between his words and the facts, and it would be pointless to do this unless one supposed that there was a discrepancy between them.

My suggestion that the act of making a statement can have pragmatic implications, which we have seen to be consonant with ordinary language, may give the impression that this sort of implication is not a logical matter at all. I have tried to show that although pragmatic inferences are not of the kind familiar in the textbooks of logic, they are nevertheless made in accordance with certain rules, and to that extent have a "logic." In conclusion I want to show that

this is confirmed by the fact that there is an analogy between pragmatic implication and strict implication or entailment.

If a man fails to see that "ABC is a triangle" strictly implies "ABC has three sides," we would account for this by saying either that he has no intelligence whatever, or that he does not know what the words "side" and "triangle" mean. Similarly, if someone should say "I know she never married, but what was her maiden name?" we would explain the question by saying that the speaker is either demented or does not know the meanings of either or both the expressions "married" and "maiden name." In both cases not knowing the meaning of an expression is the same as being unable to use it correctly. The important point here is that when we account for linguistic aberrations such as the question just cited, we do not abandon the conception of a standard use of language, that is, the idea that there is a correlation between the propositional attitudes of the speaker and his utterances—as I have argued, this would be tantamount to denying that the utterance was a statement at all. Rather, we retain the principle and account for deviations from it as the result of mental weakness or ignorance of the language. There is thus a resemblance between strict and pragmatic implication in that both are dependent upon the meanings of the terms employed, i.e. the rules governing their use.

This can be illustrated by reference to a well-known change in pragmatic implication. Hamlet's statement "I'll make a ghost of him that lets me" would now be taken to imply a desire to be forcibly restrained, if it could be taken to imply anything at all. Because Shakespeare and his contemporaries used "let" to mean "hinder," for them the statement expressed a threat and implied a vehement desire not to be restrained. We would normally explain such a change in pragmatic implication by saying that the meaning of the word "let" has been altered. But it would be *theoretically* possible to account for this change in implication and at the same time retain the original Shakespearean meaning of the word "let" by saying that nowadays when people want to be restrained they imply a desire not to be restrained. This, although fantastic, would presumably be a customary way of using language if it were usual for people to help each other as little as possible; in that case, in order to persuade another person to do something that I wished, I would use language implying that I desired the opposite, so that in his desire to displease me he would fall in with my wishes. Here the relation between propositional attitudes and language breaks down, or rather is replaced by another. Of course, changes in pragmatic implication are not explained in this bizarre way; we ascribe them to changes in meaning, that is, definition. Here again we can see a connection between pragmatic and strict implication. The logical character of propositions

may alter; p may once have entailed q although it no longer does so, and *vice versa*. We do not ascribe this to a change in the laws of logic but to a change in meaning or definition; definitions are not sacred and are liable to revision, in which case the relevant propositions will occupy a different position on the logical map. If in respect of the word "let" a distinction is drawn between Elizabethan and Georgian English, we could describe our current language either as an unreasonable usage of Elizabethan English or as a reasonable usage of Georgian English. We do not describe changes in pragmatic implication in the former way, because this would involve the postulation of new relations between propositional attitudes and utterances, and a consequent re-definition of "statement." One can describe how this would be done; it would be more difficult, though not impossible, to imagine the kinds of states of affairs which would induce us to undertake the task.

University of Nottingham.

ON SELF-REFERENCE

by W. D. Hart

THERE seems to be an at least initial plausibility in distinguishing between good and bad cases of self-reference. For instance, on one reading the sentence

(*a*) This sentence is false

is false if it is true and true if it is false. So when (*a*) is read self-referentially, it issues in paradox; thus (*a*) could be counted as a bad case of self-reference. But in contrast with (*a*), on no reading does the sentence

(*b*) This sentence is in English

seem to issue in any paradox analogous to that which (*a*) yields. Moreover, even though on one reading (*b*) is self-referential, on that same reading it also seems to be straightforwardly true; thus (*b*) could be counted as a good case of self-reference. (See, for example, Robert L. Martin, "Toward a Solution to the Liar Paradox," *Philosophical Review*, LXXVI [1967], 279-311.) There seem, however, to be certain puzzles associated with (*b*) and similar sentences.

How (if at all) should (*b*) be translated into German? Proceeding roughly word by word, a natural candidate for a translation of (*b*) into German might be

(*c*) *Dieser Satz ist auf english.*

There are two reasons for denying that (*c*) is a correct translation of (*b*) into German. First, while (*b*) seems to be true, (*c*) seems to be false; but translation should preserve truth value. Second, both (*b*) and (*c*) are *self-referential*. The grammatical subject of (*b*) refers to (*b*) and the grammatical subject of (*c*) refers to (*c*). But (*b*) and (*c*) are different sentences (since, for instance, they are in different languages). Hence, (*b*) and (*c*) refer to different things; but translation should preserve reference.

Should we wish translation to preserve truth value and self-reference (but not reference), we might take

(*d*) *Dieser Satz ist auf deutsch*

as a translation of (*b*) into German. Both (*b*) and (*d*) seem true. Moreover, intuitively they seem to be true for *almost* exactly the same reason; that is, (*b*) says of itself that it is in English and indeed it is in English, while (*d*) says of itself that it is in German and indeed it is in

German. Again, both (*b*) and (*d*) are self-referential; the grammatical subject of (*b*) refers to (*b*) and the grammatical subject of (*d*) refers to (*d*). For precisely this reason, however, if (*d*) is a translation of (*b*) into German, then translation does not always preserve reference. Moreover, intuitively it would seem that (*b*) and (*d*) ascribe *different* properties to their respective and *different* subjects; thus, if (*d*) is a translation of (*b*) into German, translations may not be synonymous.

Should we wish translation to preserve both truth value and reference, we might adopt the following device. Let us christen the sentence (*b*) "Sentence (*b*)." Now consider the sentence

(*e*) Sentence (*b*) is in English.

Both (*b*) and (*e*) seem true. Moreover, both the grammatical subject of (*b*) and the grammatical subject of (*e*) refer to sentence (*b*). Again, since both (*b*) and (*e*) say of (*b*) that it is in English, and indeed (*b*) is in English, (*b*) and (*e*) could be said to be true for exactly the same reason. Now consider the sentence

(*f*) *Der Satz* (*b*) *ist auf english.*

Then (*e*) and (*f*) seem both to be true, both to refer to (*b*), and to be true for exactly the same reason. So we might take (*f*) to be a translation of (*e*) into German. Then the relations lately noted between (*b*) and (*e*) might be sufficient for us to take (*f*) to be a translation of (*b*) into German as well.

Two points should be noted here, however, First, since the grammatical subject of (*f*) refers to (*b*), and since (*b*) and (*f*) are different sentences, and since the grammatical subject of (*f*) is a singular term, (*f*) is not self-referential. But this is as it should be; for we have lately learned that reference is invariant under translation if and only if self-reference is not. Second, if (*f*) translates *both* (*e*) and (*b*) into German (and it would seem that it will not do to take [*f*] as a translation of [*b*] unless we also take it as a translation of [*e*]), then what is the relation between (*b*) and (*e*)? It would seem that both (*b*) and (*e*) should be translations of (*f*) into English. But then either one sentence can have non-synonymous translations, or (*b*) and (*e*) are synonymous. And if (*b*) and (*e*) are synonymous, then either they are identical (which is preposterous), or self-reference is not invariant even under synonymies within a single language.

Like (*b*), the sentence

(*g*) This sentence contains five words

might seem straightforwardly true. If, however, we attempt to encode

this appearance via the Tarski Paradigm, we again encounter paradox. For the sentence

(*h*) The English sentence "This sentence contains five words" is true if and only if this sentence contains five words

seems false. Since (*g*) seems true, the first limb of biconditional (*h*) should be true. Yet the unquoted occurrence of the phrase "this sentence" in the second limb of (*h*) seems to refer to *all* of (*h*). But then the second limb of (*h*) is false, since (*h*) contains fourteen occurrences of twelve words. Hence (*h*) is false.

The difficulty with (*h*) is not to be blamed on "true" so much as on "if and only if." The sentence

(*i*) Neutrinos have mass or this sentence contains five words

seems to *follow* from the apparently *true* sentence (*g*) and yet also seems to be false for much the same reason as (*h*) is false: neutrinos have no mass and sentence (*i*) contains nine words. Nor is the difficulty confined to extensional sentence connectives; though (*g*) seems true, if the sentence

(*j*) George believes that this sentence contains five words.

is true, then George may be in error, since (*j*) contains eight words.

The principle responsible for these difficulties may be stated as follows. Concatenation and generally syntactical manipulation of sentences preserves reference if and only if it does not preserve self-reference (assuming the output of the manipulation is not identical with the input, which should be required of nontrivial manipulation). Thus, (*h*) is true only if its second, unquoted occurrence of "this sentence" is no longer self-referential; and it is only if the occurrence of "this sentence" in (*i*) is no longer self-referential that (*i*) follows from (*g*). The device suggested above to preserve reference under translation will serve equally well to preserve reference (but not self-reference) under syntactical manipulation.

Like (*b*) and (*g*), the sentence

(*k*) The sixth letter in the alphabet does not occur in this sentence

might seem straightforwardly true. But

(*l*) The sixth letter in the alphabet = the sixth letter of the alphabet

is as trivial as truths get, while the sentence

(*m*) The sixth letter of the alphabet does not occur in this sentence,

though it seems to follow from (*k*) and (*l*) by substitutivity of identity, is nonetheless false. The principle underlying this difficulty is that substitution of one term for a different term in a sentence preserves reference if and only if it does not preserve self-reference. The same device as before will serve to preserve reference.

Suppose that this paper (*n.b.*)were to be translated into German. How should it be translated? I would prefer that most of my sentences (including this sentence [*n.b.*]) be translated roughly term by term But in order that my arguments retain their validity and my examples their points, I should also prefer that (*b*) be replaced by (*d*) and vice versa, and that (*c*) be replaced by "This sentence is in German." In general, I should prefer that the term "English" be replaced by the term "*deutsch*" and vice versa, and that the term "German" be replaced by the term "*english*" and vice versa at all their occurrences including their quoted occurrences in this sentence (*n.b.*). It makes all the difference in the world to this paper in which language it is written.

There seem to be two morals to be derived from the above paragraph. First, the units, as it were, of translation are not always isolated words or sentences, but sometimes whole bodies of discourse. That is, relations between sentences can determine the appropriate correlate of an individual sentence under translation of an integrated body of discourse in which those sentences occur. What should count as the correct translation of an individual sentence may depend on whether that sentence is translated by itself or as a member of a body of related sentences. Thus, if (*b*) is translated by itself, (*f*) seems preferable to (*d*); but if (*b*) is translated in the course of translating this paper, (*d*) seems preferable to (*f*).

Second, the above paragraph but one brings out a rationale for something quite like self-reference. For in that paragraph we went some way toward considering the application of this present discussion of self-reference to itself. And we had an obligation to do so. It would seem that a theory T of theories should apply to T itself, and that an account A of our use of our words should apply to our use of words in giving account A. If, for example, T were obviously false of T itself, then even if it were true of all other theories, it would still be unsatisfactory. Generally, a variety of *argumentum ad hominem* is valid, and that variety requires self-applicability. (See F. B. Fitch, "Self-Reference in Philosophy," *Mind*, LV [1946], 64-73.)

But since self-applicability is obviously quite like self-reference, it may not be quite clear whether we can retain the benefits of self-

applicability (intuitively, a desirable sort of completeness) while avoiding the paradoxes of self-reference. Consider first the Liar.

A theory T of phenomena P (for example, theories, products of human effort, factors increasing understanding, formalizable languages) should apply to *all* P's, even if T is itself a P. More precisely (but less generally), if T is a theory of P's, if T is itself a P and if $\vdash_T (x)Q(x)$, then T should contain a term "t" which refers to T such that $\vdash_T Q(t)$. What prevents the Liar Paradox is that for T to be consistent, "sentence of T which is true" cannot be one of the predicates of T; (this is an informal version of Tarski's Theorem). T should say that $Q(t)$, but it cannot say that "$Q(t)$" is a sentence of T which is true. In order to have self-applicability, we permit self-reference; we then look to a theory of truth rather than a prohibition of self-reference to forestall the Liar. (We must therefore be prepared to except that theory of truth from our general requirement of self-applicability. But generally, we want self-reference to be possible; and, generally, a form of it will be. Let L be any countable, first-order recursively axiomatized formal system. Then L can be gödelized; that is, under an assignment of natural numbers to the expressions of L, many metalinguistic statements about the syntax of L are equivalent to intuitive arithmetic statements. If all recursive functions are numeralwise axpressible in L [and this seems to be a rather weak constraint on formalizations of most scientific theories], then an arithmetic correlate of one of the aforementioned syntactical statements about L will be true only if a certain effectively specifiable sentence of L is provable in L, and it will be false only if the negation of that sentence of L is provable in L. By this somewhat roundabout route, some statements about the syntax of L are true only if certain sentences are provable in L. Such sentences of L are thus often thought of as providing a means of "describing" some of the syntax of L in L. In this sense, gödelization makes possible a version of self-reference. But almost all actually studied formalisms can be gödelized. Hence it would be difficult, if not crippling, to attempt to prohibit all versions of self-reference. [Moreover, if L can encode some set theory, then it can "describe" some of its own model theory as well.])

Consider now the paradoxes associated with (b), (g), (k). Briefly, our solution to those paradoxes consisted in retaining (b), (g), (k), and thus permitting self-reference, while insulating (b), (g), (k) themselves from such communal linguistic activities as translation and syntactical manipulation; in the hurly-burly of interaction, (b) would be represented by its proxy (f). This general solution insures that reference

will be preserved in the give and take of communal linguistic life. But this general solution finds an exception when we are dealing with (for example, translating) an extended body of discourse, argument, and examples on the topic of self-reference itself, like this paper. Then we should let a charitable desire to preserve the validity of arguments and the points of examples and so on be our guide. The foregoing exception is but another mode of insulating self-reference in accord with the main principle. That is, we should insulate self-reference so as to preserve other more desirable phenomena such as reference, validity, purpose, and so forth; but insulating self-reference is not abandoning it, and for the sake of preserving such desirable phenomena as self-applicability, self-reference should not be abandoned.

W. D. HART

The University of Michigan

PART III

METALOGICAL SELF-REFERENCE

SELF-REFERENCE IN PHILOSOPHY.

By Frederic B. Fitch.

A THEORY always has a particular subject-matter associated with it. We say that the theory is " about " its subject-matter. For example, Darwin's theory of natural selection is about living organisms, and species of living organisms, and genetic relationships among such species. Newton's theory of universal gravitation is about particles of matter and about certain relationships of attraction between such particles. In so far as a theory is vague, the exact extent of its subject-matter tends to be hard to specify. A precisely stated theory, on the other hand, tends to have a clearly delineated subject-matter. We may ordinarily regard the subject-matter of a theory as consisting of some class of entities, together with certain subclasses of that class and certain relations among its members. The notion of " subject-matter " could be more carefully analyzed if use were made of symbolic logic, but this concept should be clear enough in the light of the informal examples just given.

Some theories are about theories. Others are not. Theories which do not include theories in their subject-matter will be said to be of *ordinal level zero.* A theory which includes in its subject-matter some theories of ordinal level zero, but none of higher ordinal level, will be said to be of *ordinal level one.* And so on. In general: A theory of ordinal level $n + 1$ includes in its subject-matter no theories of ordinal level greater than n, but it does include some of ordinal level n. Here n may be thought of as any finite or infinite ordinal number. Many theories proposed in the empirical sciences can be seen to be of some fairly low finite ordinal level. This is because empirical science is not generally concerned with framing theories about all theories.

A different situation prevails in philosophical research. Here extreme comprehensiveness is sought for. Theories are constructed which purport to deal with all entities whatsoever and which therefore have an unrestrictedly extensive subject-matter. In dealing with all entities, such theories in particular deal with all theories, since theories are themselves entities of a special sort. In philosophy we thus encounter theories about

the general nature of theories. If a theory has an ordinal level, its ordinal level must be greater than the ordinal levels of all theories occurring within its subject-matter. Hence a theory about the general nature of theories can have no ordinal level, for its ordinal level would have to be greater than itself. Theories having no ordinal level will be said to be " vertical " or " non-ordinal " theories. Theories having ordinal levels will be said to be " horizontal " or " ordinal " theories.

If a theory is included in its own subject-matter, we say that it is a *self-referential* theory. Since no ordinal level can be assigned to a self-referential theory, every self-referential theory is vertical and non-ordinal. The converse, however, is not true, because a theory might contain vertical theories in its subject-matter without containing itself in its subject-matter. Such a theory would be vertical but not self-referential.

An example of a vertical and self-referential theory is Whitehead's philosophical system as presented in *Process and Reality*. Among the entities considered in his system are not only actual occasions, eternal objects, prehensions, nexus, contrasts, and multiplicities, but also propositions or theories. His whole doctrine of these entities is itself a theory. Since it is a theory about all theories, it includes itself in its subject-matter.

Whitehead's identification of propositions with theories raises the question as to whether theories should be treated as classes of propositions or as individual propositions. Either view seems equally tenable. For present purposes Whitehead's view will be accepted. This means that every proposition is a theory, and conversely, so that every proposition is regarded as having a subject-matter. In the terminology of symbolic logic one might say that the subject-matter of a proposition consists of all entities which are values for variables occurring in the statement of the proposition, together with all entities denoted by expressions which occur as proper parts of the statement.

Some writers attempt to abandon the notion of " proposition " altogether, and to replace it by the notion of " statement " or " sentence," regarding the latter as a mere string of symbols. Such a procedure is useful as a matter of method in the field of syntax, where the meanings of symbols are not of interest so much as the symbols themselves. When questions of meaning are raised, however, this sort of nominalism seems very inadequate. Even Carnap, who formerly advocated some such nominalism, has now largely relinquished it.

Any system of philosophy which takes a position on the nature of theories or propositions is itself a vertical self-referential theory. Particular views as to what constitutes a valid or acceptable theory are also themselves vertical self-referential theories. For example, consider the view that every valid theory must be obtained from observed empirical data. This is a theory about theories and their validity. Incidentally it is a theory which does not seem to conform to its own criterion as to what constitutes a valid theory, at least not unless it can itself be shown to have been obtained as a generalization from observed empirical data. A vertical theory is always open to just this sort of danger. It may not itself conform to some principle that it lays down concerning theories in general. A horizontal theory, on the other hand, is open to no such danger. It may be internally inconsistent, or it may be inconsistent with known facts, and hence " externally " inconsistent, but it cannot be inconsistent with its own nature in the way that a self-referential theory can. If a self-referential theory T implies that T has the property P, and if T in fact does not have the property P, then we shall call T self-referentially inconsistent.

Self-referential inconsistency is important in at least two respects. In the first place, a standard method for attempting to refute a philosophical view is to show that it is self-referentially inconsistent. This is a method which can be applied only to vertical, or at least self-referential, theories. Hence it is a method which is for the most part peculiar to philosophy and philosophical logic. In the second place, self-referential inconsistency, or something almost the same, is at the heart of many important problems in logic and mathematics. Some of the most interesting problems of modern logic centre around the paradoxes of set-theory and the closely analogous semantical paradoxes. All these paradoxes involve propositions which refer to themselves or to some part of themselves. Any system of mathematics or logic in which such paradoxes can arise is both vertical and inconsistent, though it might not be actually self-referential itself. The vertical or non-ordinal aspect would arise from the fact that self-referential propositions would be part of its subject-matter. There exist restricted vertical systems of logic and mathematics which seem to be free from the paradoxes of set-theory, though consistency has not yet been definitely established in the case of the most important and useful of such systems. Even within such restricted systems it is possible to prove certain fundamental theorems due to Cantor and Gödel which are closely similar to the paradoxes of

set-theory both with respect to the presence of something analogous to self-reference (or even self-reference itself, in the case of Cantor's theorem) and with respect to the rôle played by the concept of negation. More will be said about these mathematical matters later. First let us consider the importance of the notion of self-reference in philosophical methodology.

It may be that nobody has ever seriously proposed or tried to defend a system of philosophy which was actually self-referentially inconsistent, though many systems of philosophy superficially *seem* (to those attacking such systems) to be self-referentially inconsistent. For example, consider the sceptical point of view according to which nothing is " absolutely " true. This view casts some element of doubt on every proposition. According to it no proposition can be asserted as true for certain. All theories are open to some doubt, it holds. But this view is itself a theory about all theories, and the doubt it casts on all theories it casts equally well upon itself. If it is really a valid theory, then it is wrongly questioning its own validity, in questioning the validity of all theories. Therefore if it is valid it is self-referentially inconsistent, and hence not valid after all. Therefore it cannot be valid. A similar situation is to be found in Descartes' method of doubt. He could not doubt that he was doubting, and hence he found something indubitable. Complete doubt of everything led to a self-referentially inconsistent view and so had to be abandoned. We thus get the positive result that some propositions may be affirmed with certainty. In fact we can conclude that doubt " presupposes " certainty.

The notion of " presupposition " suggests various sorts of philosophical idealism and related types of philosophy. In such philosophies a " presupposition " often seems to mean some hypothesis that cannot be systematically denied without in some sense being already assumed. The very denial itself, or some important aspect of it, or some assumption or method involved in presenting and defending it, constitutes an exception to the denial. A presupposition might be defined as an assumption whose denial is self-referentially inconsistent. For example, any systematic consideration of and rejection of the accepted principles of logic already involves the use of at least some of those principles of logic. Hence it is a presupposition that at least some of the principles of logic are valid. Similarly, any attempt to reduce the principles of logic to mere conventions regarding the use of symbols must already employ those principles themselves in carrying out the reduction. Hence the

reduction is really a reduction of logic to *conventions-plus-logic*, and logic is not completely "analyzed away" into something else.

The concept of presupposition may also be considered in connection with the theory of value. This is because value judgments enter into the theory of value, or rather into specific theories of value, not only as part of the subject-matter but also as part of the intellectual apparatus used for defending or attacking particular theses concerning value. For example, one value theorist might attack the scientific or philosophical methodology of another value theorist as "bad" or "unsound" methodology. But the attacking theorist might be assuming a theory of value according to which the phrase, "X is bad", should always be replaced by a phrase of the form, "Y dislikes X", and nevertheless might be unwilling to restate his attack in the form of a mere statement of personal dislike. If so, the attack becomes self-referentially inconsistent, inasmuch as it is based on a theory to which it does not itself conform. The rejection of the demand that phrases of the form, "X is bad", be restated in the form, "Y dislikes X", is a presupposition of every theory which makes value assumptions about ("good" and "bad") methodology and fails to treat such value assumptions as mere matters of personal like or dislike.

The type of argument in which one accuses one's opponent of self-referential inconsistency is really a very ancient type of argument. It has often been called the *ad hominem* type of argument, since it may involve the pointing out of some fact about the opponent himself which contradicts or is an exception to the view he propounds. It is perhaps best understood as a request that the opponent clarify his position sufficiently to destroy some superficial appearance of self-referential inconsistency. A solipsist, for example, might be expected to hold the view that his solipsism needs no defence against the attack of an opponent, since the solipsist maintains that nobody else, and hence no opponent, exists. Thus solipsism seems to presuppose the existence of other minds insofar as the solipsist takes the trouble to reply to objections to his view. But this is perhaps a superficial interpretation of solipsism, and a careful solipsist might state his position in such a way that it would be evident that he was stating his position for the benefit of no other mind but his own.

The *ad hominem* type of argument is probably more liable to stir up the resentment of an opponent than any other type of argument. This is because it has the appearance of being directed

at the opponent himself, as well as against his thesis. It may therefore be treated as if it were a personal insult of some sort, involving even ridicule and irony. The opponent is made to look like very much of a fool when the *ad hominem* argument is well presented, because the exception to the opponent's view is found to exist not in some distant situation but, of all places, in some situation or fact immediately involving the opponent himself. Not only does self-referential inconsistency involve a definite sort of irony, but consideration seems to reveal that all cases of irony, conversely, have in them some element of self-referential inconsistency, or something approximating to it.

The personal aspect of the *ad hominem* type of argument tends to cause it to be regarded as an " unfair " type of argument, and indeed unsound. The present writer, however, regards it as a very important sort of argument, and one that is perfectly valid against certain kinds of vertical theories. The mere fact that it cannot be used in connection with horizontal theories arising in the special sciences does not mean that it can have no application in philosophy. On the contrary, the possibility of using it in philosophical speculation and in the criticism of systems of philosophy is a mark which distinguishes philosophy from the empirical sciences. W. M. Urban in his book, *The Intelligible World*, makes repeated use of the *ad hominem* argument. On page 45 he quotes Lowes Dickinson as holding that in ultimate matters the *argumentum ad hominem* is " the only argument possible and, indeed, the only one in which anyone much believes ".

Although no *ordinal* level can be assigned to a theory which is about all theories, still we may speak of its " level " in some broader sense. A theory about all theories may be said to have attained the level of maximum theoretical generality. At such a level all other levels may be dealt with. There is no level which is higher in the sense that it can deal with theories not dealt with on the level of maximum theoretical generality. To deny that there is such a level is already to be proposing a theory about all theories and hence to be presenting a theory which is itself of the level of maximum theoretical generality. Thus an *ad hominem* argument can be used against the contention that no such level is to be found. It is characteristic of philosophy to reach this maximum level and to be able to use the self-referential sorts of reasoning which are possible on this level.

An analogous situation is to be found in the classical theory of real numbers. The real numbers can be defined as classes of rational numbers. We thus obtain numbers (namely, the

irrational real numbers) having various properties not possessed by the rational numbers. If we attempt to go a step further and define some other sort of number in terms of classes of real numbers in exactly the same way that the real numbers are defined as classes of rational numbers, then nothing essentially new or different is obtained. This is because the class of real numbers has a sort of "level of maximum numerical generality", just as a theory about all theories has a level of maximum theoretical generality. The analogy can be seen from the fact that on the classical theory of real numbers it is permissible for an individual real number to be defined in terms of the class of all real numbers. This is similar to the situation where we have a theory dealing with all theories. On the classical theory of real numbers, generally speaking, it is permissible for an entity to be defined in terms of a class (*e.g.*, the class of real numbers) having that entity as a member. Such a definition is not "circular" in the objectionable sense of defining an entity in terms of itself, but it is nevertheless circular in a secondary sense, since a class having the definiendum as a member is a factor in the definiens. Real numbers defined in terms of the class of all real numbers are thus circularly defined (in a secondary sense of "circularity") and involve self-reference. Cantor's proof that the class of real numbers cannot be put into a one-to-one correspondence with the class of rational numbers consists in supposing that the correspondence has been set up and then in defining *in terms of the correspondence* (and hence in terms of the whole class of real numbers) a particular real number that must have been omitted from the correspondence. The particular real number, of course, involves a sort of self-reference. Russell's "branched" or "ramified" theory of types of the first edition of *Principia Mathematica* was designed to do away with all self-reference in logic and mathematics in order to provide protection against the paradoxes of set-theory and the paradoxes of semantics, since Russell believed these paradoxes to be due to a "vicious" circularity. Russell proposed the Axiom of Reducibility, however, as a device to moderate (in effect, if not in theory) the elimination of all circularity and to permit the sort of secondary circularity required for Cantor's theorem. [A similar effect is obtained more simply by replacing the branched theory of types by the "simplified theory of types". This method, however, can be safely used only where semantical concepts are not being assigned type.] Unless some appropriate sort of circularity and self-reference is allowed, Cantor's theorem no longer holds and

the real numbers no longer represent a genuine maximum level. In order to get enough real numbers for mathematical purposes without some such circularity, it becomes necessary to keep proceeding to higher and higher levels (or " orders ") without ever reaching a final level on which all the real numbers may be handled. For this reason the branched theory of types, unless moderated by the required reducibility principle [or, equivalently, transformed into the simplified theory of types], is not held in esteem by most mathematicians. Something very much like the branched (" ramified ") theory of types, not too much moderated by a reducibility principle, seems nevertheless essential for avoiding the paradoxes of semantics in those theories which are concerned with *semantical* as well as mathematical concepts. The ramified theory of types, however, cannot be taken as laying down ultimate restrictions which eliminate all sorts of self-reference whatsoever. Not only would the theory of real numbers be crippled, but all theories about the totality of theories would be eliminated. Furthermore, such a ramified theory of types could not even be stated. Its sweeping restrictions against self-reference would apply to every theory, including itself, and so it would be self-referential in violation of its own edicts. A similar criticism can be made even against the more moderate simplified theory of types, if regarded as universally applicable. This sort of criticism is clearly just another instance of a use of the *ad hominem* argument. One way of attempting to meet this objection to the ramified or simplified theory of types is to assert that a formulation of a theory of types is simply the formulation of certain more or less arbitrary or conventional stipulations about the permitted ways of combining symbols. This answer seems to be all right so long as one is restricting oneself to the realm of uninterpreted symbols, but as soon as one enters the realm of semantical concepts it becomes necessary to apply distinctions of " type " to *meanings* of symbols as well as to symbols themselves, and the element of self-reference reappears. For example, the ramified theory of types cannot assign a type to the meaning of the word " type ", and yet it must do so if the theory applies to all meanings. In a similar way, no " order " (in the sense used in the ramified theory of types) can be assigned to a proposition which is about all propositions, hence no order can be assigned to the proposition which states the ramified theory of types.

The problem is to find a theory of types which eliminates the " vicious " sorts of self-reference that lead to the mathematical

and semantical paradoxes but not those sorts of self-reference that seem to be such an important part of philosophical logic, or required in developing the theory of real numbers.

One way of constructing such a theory of types might be somewhat as follows: We should first relinquish the view that to every proposition corresponds another proposition which is its "contradictory" or "denial". Only certain explicitly specified propositions, or propositions of certain explicitly specified sorts, would be regarded as possessing denials. To say that a proposition p is "false" would be to say that p has a denial and that the denial of p is true. A true proposition may possess a denial, but it cannot possess a true denial. The denial of a true proposition may perhaps be regarded as an unrealized possible state of affairs, or if the true proposition is *necessarily* true, then as an impossible state of affairs.

Propositions of the following kinds could be treated as possessing denials:

(1) Those to which an ordinal level is assignable, and in particular those which conform to the ramified (branched) theory of types, since there is no danger of paradoxes arising in connection with such propositions.

(2) Those which involve no direct or indirect use of the concept of denial (negation), since these propositions are free from the danger of producing paradoxes. [The inclusion of this category of propositions was suggested to me by Mrs. Gladys Barry. A large and important part of logic can be dealt with by means of propositions which involve no use of negation. See *The Journal of Symbolic Logic*, vol. 9, 1944, pp. 89-94.]

(3) Propositions required for stating that various propositions do or do not have denials.

(4) Propositions conforming to the simplified theory of types and not involving any semantical concepts. [This category of propositions could be omitted if the axiom of reducibility is employed in connection with the ramified theory of types].

An example of a proposition belonging to category (3) above would be the following: "All propositions have denials". This is regarded as a false self-referential proposition. It has a denial and its denial is true.

Here is an example of a proposition which does not belong to any of the categories (1)-(4) above and which is therefore regarded as not possessing a denial: *Every italicized proposition appearing in this paragraph is false.* This italicized proposition implicitly says of itself that it has a denial and that its denial is true. If it did have a denial, it would be an example of a

proposition equivalent to its own denial, and this situation would constitute an unacceptable contradiction. This contradiction is of course at once recognized as a form of the Epimenides paradox. The method used here for avoiding it seems to apply equally well to the other standard logical paradoxes.

UNIVERSAL METALANGUAGES FOR PHILOSOPHY

FREDERIC B. FITCH

ONE WAY to try to be clear in philosophy is to formalize, or try to formalize, the basic principles of one's system of philosophy. The mere process of attempting to translate philosophical statements into the language of symbolic logic forces one to think out the details and content of philosophical concepts with an unusual amount of precision. The task of formalizing philosophy seems to require at least three kinds of special ability:

(a) Philosophical ability, so that the principles chosen for formalization are not trivial or absurd.

(b) Ability to perform natural language analysis preparatory to transforming natural language statements of philosophy into statements of a formal language.

(c) Ability to construct a suitable formal language to serve as the vehicle of the formalization.

Needless to say, there have to date been very few, if any, successful formalizations of philosophical principles. The difficulty has so far proved to be too great, though the ideal has often been recognized as one worth seeking to achieve. My present purpose, however, is to give some indication that the task of formalizing philosophy is actually not quite so difficult or impossible as certain considerations have in the past made it seem, and that ways are open that have not yet been fully explored.

One reason for general pessimism regarding attempts to formalize philosophy is Tarski's famous result that the concept of truth, relative to a formal language, cannot be defined in that language itself, at least not if the language is of a certain standard kind and is consistent.[1] This theorem of Tarski seems to indicate that there is no one formal language which is adequate for all philosophical discourse. Each formal language has asso-

[1] Alfred Tarski, "Der Wahrheitsbegriff in den formalisierten Sprachen," *Studia Philosophica*, I (1936), pp. 261-405. An English translation of this by J. H. Woodger is to be found in Alfred Tarski, *Logic, Semantics, Metamathematics* (Oxford, 1956), pp. 152-278.

ciated with it a concept of truth which can be adequately represented only in a higher formal metalanguage, and that higher formal metalanguage, again, has its own concept of truth which can be adequately represented only in a still higher formal metalanguage, and so on. Thus there seems to be no final formal language adequate for dealing with its own semantical concept of truth, as well as with those of all lower languages. The various usual forms of the theory of types also lead to a similar conclusion, since we are prohibited, by each such theory of types, from making formal semantical statements about all types, and thus prohibited by each theory of types from formalizing the semantics of that theory of types itself. In all these cases we seem to be forced to resort to natural language as the medium for expressing our final generalizations about language, about types, and about semantic truth. No formal language seems able to comply with the requirements for serving as an ultimate universal metalanguage within which philosophy, or at least philosophical theories of some depth and insight, could be formalized. If we cannot find a formal language within which we can make fairly adequate statements about its own semantic truth or about semantic truth in general, how could we find a language adequate for the vast array of other philosophical concepts such as knowledge, belief, value, reality, mind, and life? The outlook does indeed seem dim in this direction, once Tarski's result has been accepted as conclusive, or once we have committed ourselves to some one of the usual forms of the theory of types. It is not hard to see, in view of these considerations, why so many philosophers have turned toward natural language and away from formal language.

As I have earlier suggested, however, matters are not really so discouraging for the formalist. In particular, I want to emphasize that Tarski's result is very far from decisive as soon as we consider some kinds of formal languages that differ in suitable ways from the formal languages that Tarski considered. The way now indeed appears to be open to construct formal languages that can adequately deal with their own concepts of semantical truth, and we are no longer forced to try to climb either an unending topless ladder of formal metalanguages, or a ladder of formal metalanguages that ends with a natural language at the top.

Instead, a formal metalanguage can be found which can provide a definition of its own truth.

Let us briefly recall Tarski's criterion T for an acceptable definition of truth. This criterion is to be used to judge the acceptability of a definition of a class *Tr* consisting of all the true sentences of an object language O, the definition of *Tr* having been expressed in a metalanguage M. The criterion requires, in essence, that the definition have as logical consequences all sentences which can be obtained from the following expression,

$$(x \in Tr) \text{ if and only if } p \qquad (1)$$

by substituting for the symbol "x" a structural-descriptive name of any sentence of the object language O, and for the symbol "p" the expression which is the translation of this sentence in the metalanguage M.

Notice that the concept of truth referred to here is a class of sentences rather than a class of propositions.

Suppose now that we use Gödel's method of representing syntax in elementary number theory and assign a so-called Gödel number to each well-formed expression of the object language.[2] We assume that the metalanguage M contains enough of elementary number theory to make this procedure feasible. We can then use names of numbers as names, also, of the corresponding expressions of the object language. Suppose, furthermore, that we identify the object language O with the metalanguage M. Tarski's criterion T then simply requires that the definition of *Tr* have as logical consequences all sentences of M that can be obtained from (1) by substituting for the symbol "x" the name of the Gödel number of a sentence of M, and for the symbol "p" that sentence itself. In this case the class *Tr* will be the class of Gödel numbers of true sentences of M, and this amounts to the same thing as saying that *Tr* is the class of true sentences of M, since we are representing sentences by their Gödel numbers. A variation of Gödel's well-known diagonal argument then shows, however, that there can be no such class as *Tr*, at least not if M is consistent. This variation

[2] Kurt Gödel, "Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I," *Monatshefte für Mathematik und Physik*, XXXVIII (1931), pp. 173-198.

of Gödel's diagonal argument amounts to formalizing in M the well-known paradox about the word "heterological," called Grelling's paradox. This is exactly how, in effect, Tarski proved that no sufficiently strong consistent formal language M could contain an acceptable definition of its own truth, and hence that some higher formal metalanguage would always be required for such a definition.

Now I wish to give evidence that Tarski's criterion T is unnecessarily restrictive, at least as applied to systems in which not all propositions satisfy the principle of excluded middle, and I wish to propose a weaker criterion T′ which is not open to the objection of being too restrictive. The criterion T′ is just like Tarski's criterion T except that in place of (1) we have (1′) as follows:

$$(x \in Tr)\, Cp \qquad (1')$$

where "C" is a sentential connective of M which is reflexive, symmetrical, and transitive and which satisfies a rule like modus ponens, that is, has the property that for all sentences "s_1" and "s_2," if "s_1" and "$s_1\ C\ s_2$" are provable in M, so is "s_2."

It is clear that under ordinary circumstances any definition of truth that satisfies Tarski's criterion T will also satisfy the criterion T′. This is because biconditionality used in (1) is ordinarily treated as a connective of the kind described in (1)′.

An example of a system or language that can define its own truth in conformity with T′ but not with T is a system CΔ which I have described elsewhere.[3] The system CΔ lacks the principle of excluded middle but nevertheless contains all of elementary number theory and a considerable part of the rest of mathematics, perhaps all that is really needed in applications of mathematics to empirical science. The connective that plays the role of the connective "C" of (1′) is the connective "$=$" of CΔ, and it may be thought of as expressing a relation of logical equivalence, or even a relation of full identity if we wish to be as extensional as possible. This relation of logical equivalence is such that some propositions in CΔ can be shown to be logically equivalent to their own denials, but

[3] Frederic B. Fitch, "The system CΔ of combinatory logic," forthcoming in the *Journal of Symbolic Logic*. This has also appeared as Technical Report No. 13, Office of Naval Research, Contract SAR/Nonr-609(16).

this result does not lead to contradiction. The reason it does not lead to contradiction is because the propositions in question fail to satisfy excluded middle. In particular, the Russell paradox gives rise to a proposition that is logically equivalent to its own denial, and so does the Grelling paradox, but in neither case does a contradiction arise. The system CΔ, in fact, can be shown to be free from contradiction. Thus by using the weaker criterion T′ we permit a definition of truth in CΔ that leads to no contradiction because the Grelling paradox itself is innocuous in CΔ, and we are no longer forced to accept an unending hierarchy of metalanguages, each of which can deal with only those languages below it in the hierarchy.

Two main conclusions that we have reached can be summarized as follows: (a) There exist systems which lack excluded middle and which can formulate satisfactory definitions of their own truth, the system CΔ being a case in point. (b) Tarski's criterion of truth is too restrictive, since an intuitively acceptable definition of truth that can be constructed in CΔ fails to satisfy Tarski's criterion T but does satisfy the different and apparently satisfactory criterion T′.

Some indication will now be given as to how the definition of truth can actually be constructed in CΔ. We can suppose that Gödel numbers are being used in the usual way to represent syntactical expressions. Before considering how the class of *true sentences* is definable in CΔ, observe that the set of *true propositions* is easily definable in CΔ. Since CΔ is a system of combinatory logic containing, as a subsystem, the standard theory of combinators, and since there is a combinator I having the property, $Ip=p$ (where "$=$" denotes logical equivalence), we may treat I as representing the class of true propositions. Thus "Ip" asserts that p is true, and we have the theorem that the proposition that p is true is logically equivalent to the proposition p itself. Thus we have a definition of the class of true propositions. We wish, however, to have a definition of the class of true sentences. Let us relate propositions to sentences by assigning to each proposition the same Gödel number that is assigned to a sentence expressing that proposition. This actually will result in assigning more than one Gödel number to a proposition, since logically equivalent

propositions are interchangeable with each other in the system CΔ, and so share the same Gödel numbers with each other. But this multiplicity of Gödel numbers will cause no trouble. Now each syntactical expression is represented by its Gödel number for purposes of dealing with that syntactical expression in CΔ. The class of Gödel numbers of true sentences therefore represents the class of true sentences, and this class of Gödel numbers is the same as the class of Gödel numbers of true propositions. There is no difficulty in defining in CΔ the relation of each proposition to its Gödel number, and since we can define the class of true propositions in CΔ, there is no difficulty in defining the class of Gödel numbers of true propositions. But this is just the class that represents the class of true sentences. If we abbreviate the name of this class in CΔ by "*Tr*" and if we choose the connective "*C*" of (1′) as the connective " = " of CΔ, then it can be shown that (1′) is satisfied by "*Tr*," where the metalanguage and object language are both chosen as CΔ.

Tarski's criterion seems to force us into an unending hierarchy of formal metalanguages because his criterion is appropriate for those languages in which Grelling's paradox, by way of a definition of truth for the language within the language itself, gives rise to inconsistency, so that it is sufficient for Tarski to invoke a formalization of Grelling's paradox in order to show the need for a higher metalanguage. By liberalizing Tarski's criterion T into a different criterion T′, provision is made for languages which avoid Grelling's paradox by lacking excluded middle and which can consequently possess a definition of their own truth. Thus the groundwork is laid for an ultimate formal system which can deal adequately with its own concept of truth, and within which various other general philosophical concepts, in addition to truth, can presumably also be formalized.

In conclusion it should be pointed out that the kind of negation used in the system CΔ is a *limited* kind of negation, limited in the sense that the principle of excluded middle does not hold with respect to it, since a proposition might fail to be true but also fail to be negated (i.e., false) is this special limited way. Suppose we use the symbol 'N_1' temporarily to designate this limited negation. In order to say that N_1 does not satisfy excluded middle, we

need a less limited, but still somewhat limited negation N_2, and in order to say that N_2 does not satisfy excluded middle, we need an even less limited, but still somewhat limited negation N_3, and so on. The reason that N_1 cannot be used to assert that N_1 fails to satisfy excluded middle is that such a statement would take the form,

$$(\exists p)\,[p \vee N_1 p],$$

and, by way of DeMorgan's Theorem, which holds in CΔ, the above statement would be equivalent to the contradictory statement,

$$(\exists p)\,[N_1 p \,\&\, p].$$

Although the system CΔ contains only the first member of this infinite hierarchy of limited negations, it seems likely that all the negations of this hierarchy should be included in an expanded form of the system CΔ. In the resulting system it would be possible to say of each negation that it failed to satisfy excluded middle, since a less limited negation, higher up in the hierarchy, could always be used to make such a statement about a lower or more limited negation. Thus instead of a hierarchy of formal metalanguages we may be forced to accept a hierarchy of limited negations, at least if we wish to say, in a formal way, some of the things about these various negations, and in particular about N_1, that appear to be true. It might therefore seem that we are no better off after all, since a hierarchy of negations is in some ways very much like a hierarchy of formal metalanguages. But there is one important difference. We can make statements about all the various limited negations in a single universal formal metalanguage, but there is no single universal formal metalanguage in which we can make statements about all the members of the hierarchy of formal metalanguages. Thus the hierarchy of limited negations appears to be a substantial improvement over the hierarchy of formal metalanguages.

In any case, the situation of a formalistic philosopher, who wishes to formalize his basic philosophical tenets in a universal formal metalanguage, is much more promising than Tarski's somewhat restrictive criterion of truth might at first lead one to believe.

Yale University.

THE IDEA OF A METALOGIC OF REFERENCE *

by

STEVEN BARTLETT
Department of Philosophy, Saint Louis University, Saint Louis, Missouri

Introduction

I would like to address the interests of an approach in philosophy which seeks to disclose and to investigate basic commitments involved when questions are raised about the possibility of experience, the possibility of knowledge, or the possibility of theory in general. A concern for the structure of the possible has, since Kant, traditionally gone by the name 'transcendental'. The basic commitments or investments involved in doing transcendental philosophy will be central to what I wish to treat here. In a sense, then, the context for what follows intends to offer a basis for a metacritique of transcendental philosophy.

I have been persuaded that a transcendental approach can gain a helpful measure of clarity and precision by shifting from the traditional Kantian perspective to a point of view that emphasizes the nature of referring. This shift, as I propose to describe it, provides an effective means for confirming transcendental results. A need for ways to demonstrate the validity of transcendental claims will bring me to a discussion of what I term 'self-validating logics'.

To be specific, I will (i) suggest a rationale behind shifting to the perspective of referring, (ii) propose a general *metalogic of reference* that retains the interests of transcendental philosophy, (iii) describe the usefulness of self-validating logics in this context, and (iv) conclude with some remarks about the value of transcendental philosophy, referring, and the idea of self-validating logics for philosophy of science.

* Research reported here was partially supported by a grant from the Max-Planck-Gesellschaft.

The transcendental approach

Transcendental philosophy finds its orientation in a movement away from a reflection on the actual as such, to a study of the preconditions of its *possibility*. The concept of possibility is fundamental to the transcendental approach, although exactly what possibility *is* has remained vague in the literature of the transcendental perspective. There are a number of alternative conceptions of possibility. I will suggest six of these, in an approximate order of concepts of increasing generality. This sequence will serve to determine a highly general, comprehensive sense of possibility, in terms of which a rationale for the shift I propose will be evident. The alternative views are these:

1. What is possible refers to future alternative states of a physical system.
2. The Stoic-Diodorean view: What is possible refers to what is or will be.
   The Aristotelian-Megarian view: What is possible includes what is, will be, or has been.
3. What is possible relates to the status of a description of an event which is *not excluded* by the known laws of nature.
4. What is possible is classically free of contradiction.
5. What is possible includes those real or abstract objects of reference, of which we can predicate what are ordinarily considered to be incompatible propositions.
   (I have in mind such ascriptions of properties as are frequently termed 'complementary' in elementary particle physics).

To these five views of possibility, a sixth is added that offers some promise as a highly inclusive concept of possibility.

6. What is possible can be understood as a function of an analysis of preconditions of *valid referring*. This view will be developed in what follows.

The general idea of a metalogic of reference

It is convenient to talk about referring in the context of an analysis of descriptions. Both in ordinary usage and in the natural and behavioral

sciences descriptions are relied upon to serve a variety of referential functions.

The referential character of descriptions can be analyzed in terms of the commitments descriptions entail. A description presupposes certain commitments to a framework or family of similarly constituted frameworks. These commitments can be made explicit by thinking, for example, of the general, frequently quite vague, rules or conventions which lend some form of organization to admissible descriptions that can be articulated in the context of a given framework of reference. Since the relationship between conditions of reference and any description is logically prior (in the sense intended by transcendental philosophy in the Kantian tradition) to the formulation of any specific description as a necessary presupposition of it, it seems justified to speak of "referential preconditions".

Referential preconditions are restrictions The hierarchy of different concepts of possibility 1. - 6. is actually a list of various *ways of enforcing restrictions* as to what sorts of possibles we are prepared to speak of. So, an interest in preconditions of reference can be understood as an interest in sketching out a certain sort of general map of a domain of objects for which we want to assure the possibility of valid referring.

These preconditions of reference can be approached in either of two different ways: On the one hand, a study may be undertaken of a specific framework of identification: e.g., the framework presupposed in developing a general phenomenology of human visual perception, or the framework presupposed by quantum mechanical descriptions, involving the use of special kinds of measuring devices as well as an explicit or implicit theory of measurement which permits the significant use of apparatus and interpretation of observations. On the other hand, a study may be undertaken of the very general principles which seem to underlie an entire group of special identification frameworks: e.g., from the standpoint of a phenomenological account of objectivity, the group of identification frameworks - visual, auditory, tactile, etc. - that together provide a basis for the constitution of objectivity, or alternatively, the family of conceptual frameworks with which we are acquainted in the natural sciences, which together determine what is to be understood by 'nature'.

It is in this second sense - the sense in which a study is possible of the general principles of reference that underlie a group of identification

frameworks - that it is appropriate to speak of a general *metalogic of reference.* At this point, then, a metalogic of reference appears to furnish a context for a reflection on the nature of theories in general, where specific cases may be a theory of experience, a theory of knowledge, or any of the various natural or behavioral scientific theories.

Initially, then, my interest is a purely abstract one - without regard for any special theoretical identification framework; without attending, at least in the beginning, to framework-specific rules and conventions - in short, to study pervasive constraints that condition valid referring.

One approach to these highly general and abstract metalogical preconditions of referring is suggested if we think in terms of the kinds of second-order constraints which first-order constraints of a special identification framework must obey to avoid *self-referential inconsistency.*

What I mean by 'self-referential inconsistency' would involve a more technical discussion than I can undertake here, but the basic idea is simple. It is this: Paul Lorenzen, in a different context, refers to what he calls "elementary ways of speaking". He says:

> the decision to accept elementary ways of speaking is not a matter of argument. It does not make sense to ask for an 'explanation', or to ask for a 'reason'. For to ask for such things demands a much more complicated use of language than the use of elementary sentences itself. If you ask such questions, in other words, you have already accepted at least the use of elementary sentences.*

A self-referentially inconsistent use of elementary sentences in Lorenzen's context would involve the decision to employ elementary sentences in doubting the justification of using them.

The main difference between Lorenzen's view and the idea of a metalogic of reference lies in this fact: In a metalogic of reference we are concerned not with elementary usages of the language we, in fact, employ, but with "elementary" means of referring of such a kind that they immediately are involved if we consider referring as a pure possibility. In other words, the very *possibility* of calling such means of referring

* Paul Lorenzen: *Normative Logic and Ethics* (Mannheim/Zürich: Bibliographisches Institut 1969), p. 14. Cf. also P. Lorenzen: *Einführung in die operative Logik und Mathematik* (Berlin: Springer 1969).

into question presupposes them as elementary.

It is here that the idea of a metalogic of reference can be developed by resorting to what I call *self-validating logics*. I am motivated to talk about *logics* in order to furnish a context-relative means to test the same kinds of claims which a Kantian transcendental deduction seeks to justify. A self-validating logic, unlike a transcendental deduction, is fairly simple.

To make clear what I have in mind, let us suppose we wish to study what we believe to be a basic premiss of referring:

If we assume we want to think or talk about a collection of objects of various sorts, we are compelled to allow some means for this thinking or talking about them to proceed - we must be permitted somehow to refer to what we want to think or talk about. This is trivially true, and therefore I take it as basic.

Consider a candidate for a postulate in a metalogic of reference: If a metalogic of reference is to constitute a self-validating logic (or family of logics), then its axioms and postulates will themselves be self-validating, in this sense:

> *A postulate is self-validating if its denial will result in self-referential inconsistency.*

Let us consider the following as a potential elementary postulate for referring, which it seems apt to describe as a "rule of referential counter-exemplification":

> The assertion of the impossibility of referring to an individual something metalogically implies that reference is made to that thing.

This postulate *self-validates* as follows: Reference must be made to that individual something if it is to be possible to say that reference to it is impossible. The self-validation consists in the fact that a denial of the possibility of referring to an individual something is self-referentially inconsistent.

Now, if for the purposes of my informal treatment here it can be allowed that it may be possible to determine a significant number of self-validating axioms and postulates, and then to relate them in a unified and well-ordered formal system, then we would arrive at the idea of a

self-validating logic. It would differ from an ordinary formal system in that its elementary propositions are not merely postulated with some element of arbitrariness, but present themselves as compelling our assent to them if we are to be *able* to refer at all, somewhat in the manner of Lorenzen's elementary ways of speaking.

Were this to be accomplished, we would gain a significant measure of metalogical understanding of the most fundamental commitments involved in referring, an understanding that can be justified by appeal to self-validating demonstrations.

We are then not very far from being able to apply these results so as to better our understanding of, for example, the fundamental structure of a natural scientific theory. For a theory, there will be some domain(s) of objects in which its interests lie, and there will of necessity be an assortment of ways at the disposal of the scientist to refer to the objects he studies. The scientist is particularly desirous, one might add, of supplying a basis for the kind of referring his formal theory, schema of interpretation, and domain of objects oblige him to have. For, as a scientist, he chooses to respond to a need to bring the referring descriptions for which his conceptual framework provides a basis as close to the ideal of unambiguous identification as possible. And this objective is satisfiable only if fundamental commitments involved in the scientist's use of referring descriptions are made explicit and can be seen not to conflict with the theoretic claims he wishes to make.

Transcendental philosophy of science taken in this sense has several functions: to elucidate the referential preconditions basic to specific theories and shared by groups of theories, to detect self-referentially inconsistent patterns of referring, and finally to suggest valid ways of referring to replace unsound ones. To these descriptive, critical, and prescriptive functions may be added a fourth - a preventative task: to furnish guidelines in the form of a usable metalogic which can serve the interests of self-conscious and consistent theory construction.

The association with the medical model is obvious. The descriptive, critical-diagnostic, prescriptive, and preventative functions in a transcendental philosophical context are intended to contribute to the needs of theoretical soundness; the physician accepts identical functions in attending to the needs of human physical and emotional health. The

analogy between sickness and theoretical inconsistency, between medicine and philosophical therapy may to some comprise a repugnant model, but, at the same time, to deny description criticism, or criticism positive prescription, or positive prescription preventative recommendation, will to many seem arbitrary and irresponsible.

An interesting and useful philosophical reflection on the foundational structure of scientific theories I believe is offered by a metalogic of reference: Critical and close attention would be paid to the interconnection between the ways of referring essential to a theory and the objects to which the possibility of access is thereby assured, and between these ways of referring to a domain of objects, and the interpretation placed upon findings from that perspective. The understanding acquired could not wish to take the place of the natural scientist's own comprehension of his field, but it would be a qualitatively different kind of understanding, perhaps more analytically self-conscious, and ought, one would think, serve to enhance, to complement, and to render more precise the outlook of both unphilosophical scientists and non-scientifically oriented philosophers.

Summary

The author shifts the perspective of transcendental philosophy from its traditional Kantian orientation to the point of view afforded by an analysis of preconditions of referring. This shift in perspective is proposed in order to gain clarity and precision, and to provide a means for demonstrating certain of the results of transcendental philosophy.

An attempt is made to achieve systematic clarity for a concept central to the transcendental approach, the concept of possibility. The idea of a general *metalogic of reference* is proposed as supplying a highly inclusive framework from the standpoint of which preconditions of possible reference can be investigated.

The usefulness of *self-validating logics* for transcendental philosophy is suggested as furnishing a metalogical resource for transcendental demonstration.

The author concludes with a discussion of the value of a transcendental metalogic of reference for philosophy of science.

Dr Steven Bartlett was educated at the University of Santa Clara, Raymond College of the University of the Pacific, University of California, Santa Barbara, and the Université de Paris, where he received his doctorate for work under the direction of Paul Ricoeur. Dr. Bartlett was a research fellow in philosophy and mathematics at the Center for the Study of Democratic Institutions, 1969-70, has taught at the University of Florida, the University of Hartford, and is presently Associate Professor of Philosophy at Saint Louis University. He was a visiting research fellow at the Max-Planck-Institut zur Erforschung der Lebensbedingungen der wissenschaftlich-technischen Welt, Starnberg, West Germany, 1974-75. Dr. Bartlett has published a book, *VALIDITY: A Learning Game Approach to Mathematical Logic* (Lebon Press 1973), and has contributed articles and reviews to a number of journals, among them *Dialectica, etc.: a Review of General Semantics*, and *Roczniki Filozoficzne*. Dr. Bartlett is special consultant to the Rand-N.S.F. project in Regional Analysis and Management of Environmental Systems.

STEVEN J. BARTLETT

REFERENTIAL CONSISTENCY AS A CRITERION OF MEANING*

Criteria of meaning which have been proposed in the past have failed to persuade general acceptance. They have usually endorsed then-current scientific practice, or have favored the adoption of a special, usually empirical, framework. The historical failure of criteria of meaning has been due to their apparently arbitrary status as standards external to the sets of statements to which they would apply. Often, such criteria have also failed to qualify as meaningful in the test of self-application.

It is my purpose here to show that there is available to us a criterion of meaning which must be satisfied in order for individual claims, concepts, and frameworks to qualify as "meaningful". The criterion I shall recommend is that of "referential consistency". It is proposed as a criterion of meaning in the largely negative sense that non-satisfaction of the criterion involves a certain type of meaninglessness that has received little attention. The criterion developed here therefore does not express a sufficient condition of meaningfulness. One may indeed seriously doubt whether a sufficient condition can be formulated. As a result of this limitation of focus, little will be found here about the nature of meaning. On the other hand, the criterion proposed defines an important *lower limit* of meaning, below which claims, concepts, and frameworks become self-undermining. It is in this latter sense that the criterion proposed can provide a useful tool for internal analysis and criticism.[1]

The criterion I shall suggest has these rather unique properties: Acceptance of the criterion is non-arbitrary or compelling in a sense we shall explore briefly. And applications of the criterion avoid begging the question in a way in which appeals to external standards do not.

Logical criteria for evaluating, e.g, the validity of an argument, or for assessing the consistency of a theory, define for us the limits of acceptability which argumentation or theory construction endorses. To a large degree, such criteria are arbitrary in the sense that they can be changed if our purposes are served by such a change. Seen as conventions we accept in the light of our objectives,[2] the criteria

Synthese **52** (1982) 267–282. 0039–7857/82/0522–0267 $01.60

which delimit what we will accept are seldom, if ever, absolute. That is, we are not normally compelled, on pain of incoherence, to accept certain particular criteria rather than certain others, although it is often the case that, if we are to hold to our purposes, we must abide by these or related criteria if we are to accomplish what we intend.[3] In general, then, I shall call a criterion *non-arbitrary* or *compelling* if non-satisfaction of that criterion precludes achieving the task at hand. We shall look at this claim in more detail in a moment.

The criteria which define what we mean, e.g., by 'validity' and 'consistency' are "logically arbitrary" in several ways. If we detect that a criterion, or equivalently here, a rule, has been broken, we are free to amend the rule (and perhaps in so doing change the ends which the rule may serve), or correct the violation, or leave things as they are, or shift our perspective, perhaps to a more general point of view, and perceive the breaking of the rule as conforming to a more general rule in relation to which it is no longer identified as a violation. And we may have other options. But whatever the special nature of the case may be, criteria of the sort used to assess the validity of arguments and the consistency of theories constitute logically arbitrary rules for playing certain games: rules are the logical features of practical activity; the control which they make possible is a control which we choose to have, and we are free to choose otherwise.

In relation to our chosen purposes, then, logical criteria seldom compel us by reason of logic alone to accept these criteria and no others. There is, often and in general, a sense of "open-texture" about our objectives. The formal constraints we do accept may be selected because they reinforce other ends we intend: economy, comprehension, concinnity, etc. *How* we do or should make selections from among alternative, logically arbitrary criteria will not be examined here.

From the standpoint of the criteria we accept, our purposes are underdetermined or specified with a degree of vagueness to just the degree that these criteria are logically arbitrary. It is perhaps fair to say that attempts to delimit meaning by means of a necessary and sufficient criterion have failed because of this logical arbitrariness. The numerous criteria that have been recommended for detecting meaningless concepts and statements have very much the same status as do criteria which permit evaluations of validity, consistency, etc.

Criteria of meaning have come to be considered in the same game-relative light as have rules of logical evaluation.

For example, Hume, Schlick, Ayer, and Carnap have proposed these as criteria of meaning:

For Hume: expression of abstract or empirical reasoning.[4]

For Schlick: association of conditions with a proposition or question which define what experience(s) would make that proposition true, or which would if satisfied answer that question.[5]

For Ayer: verifiability, reflecting an individual's knowing how to verify a proposition which is factually significant to him.[6]

For Carnap: ability to give rules according to which observable effects can be deduced,[7] or alternately, expression of factual content.[8]

These criteria, not exhaustive of those proposed in the literature, nor yet mutually exclusive, share two characteristics: First, from a non-partisan viewpoint, it may be fair to say that acceptance of one or more of these criteria is a function of one's purposes. Second, neither Hume, nor Schlick, nor Ayer, nor Carnap, nor any other proponent of a criterion of meaning apparently has been able to show that acceptance of a certain criterion of meaning compels assent, i.e., is non-arbitrary in the sense we have sketched.

This observation would not reflect a negative judgment if, as could be claimed, we wish a criterion of meaning to function with the same measure of arbitrariness in the framework of a set of concerns as does a rule-based convention of logical evaluation.[9] But this state of affairs would clearly not satisfy authors of meaning criteria.

Criteria of meaning, then, have functioned in an *external* capacity: When they are applied, they are used to evaluate statements, concepts, or frameworks, as it were, from the outside. Criteria of meaning, understood as stipulative, normative conventions, can only be recommended in a manner which seeks to persuade our acceptance, since they do not, in and of themselves, compel assent.[10]

One of the most persuasive cases that can be made on behalf of the choice of a certain criterion of meaning is that its meaningfulness follows from its self-application.[11] If the criterion recommends that meaning be identified with expression of factual content, for example, it may be argued that 'factual content', understood in terms of operations which define the criterion, itself expresses factual content.

However, the self-applicability of a criterion of meaning, when

assured, at most insulates the use of the criterion from internal inconsistency, and may strengthen the *feeling* that its choice is not totally arbitrary. Beyond this, self-applicability does not do much: The decision to adopt a particular criterion of meaning remains external to the class(es) of statements and concepts to which it is to apply.

REFERENTIAL CONSISTENCY AS AN INTRINSICALLY DETERMINED CRITERION OF MEANING

In the view I have attempted to represent, rules for evaluating logical validity and consistency and criteria of meaning share the property of arbitrariness as game-relative conventions. The selection of such rules and criteria hence may be considered predominantly to be a function of our practical concerns. With respect to the decision to adopt a particular criterion, there is little that can be said if more than practical justification is desired. In a given field of study, rule-based evaluative conventions of one kind or another may be convenient, expedient, or necessary in practice. If one chooses to work in that field, he may have need of some externally imposed evaluative conventions. But the use of such external standards of evaluation cannot, as we have seen, be expected to be non-arbitrary and compelling.[12]

Fortunately, *there does exist a logically compelling basis for evaluation, a basis which one cannot not accept.* I have called this basis for evaluation 'referential consistency'.[13]

Referential consistency does not represent an externally imposed convention, a normative stipulation, an arbitrarily endorsed special rule or criterion. The approach to referential consistency described here rather has the character of a *metalogic*, in terms of which "preconditions of reference" in special contexts can be studied. In rough terms, initially, referential consistency is a metalogical criterion or rule of evaluation which addresses, *intrinsically*, the context-relative use of expressions, statements, or concepts. A special set of evaluative rules or criteria is not applied across the board in an external way, but rather *attention is given to those conditions which must be satisfied in a given context in order for references made in that context to be possible at all.* The results of applying such a metalogical criterion of referential consistency are non-arbitrary, both

because a special criterion is not imposed externally, and because these results compel assent – one cannot reject them in a given context of reference.

A short account of the proposed metalogic of reference will be given here. A complete formulation of the general theory will be found elsewhere,[14] as are illustrations of certain applications of the metalogic.[15]

A METALOGIC OF REFERENCE

For the sake of simplicity, I limit my treatment here to the set of *referring sentences* (alternatively, *propositions*) $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ where p_i may refer to any one or more o_i of a set of objects of reference $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$, and may possess any truth-value of a set of possible values $V = \{0, 1, \ldots, n\}$, where $n \geq 3$.[16] By the 'significant range of V' is meant '$\{0, 1, \ldots, n-1\}$'. (A discussion of the value v_n follows below.) It is clear that the significant range of V is bivalent when $n - 1 = 1$, with '0' and '1' representing the values 'false' and 'true', respectively.

Some definitions are called for.

(D1) A *particular* is a possible object of identifying reference.

Alternatively,

(D2) An *identifying reference* is such that an ascription to that which can be the subject of an ascription (namely, a particular) establishes that what is ascribed (one or more properties, relations, a description, etc.) and that that to which the ascription is made are one and the same (identification).

D1 contains the there undefined concept of identifying reference, while D2 leaves undefined the concepts of particular, description, property, relation, identification, and ascription.

In the interests of simplicity we will retain D1, permitting the concept of identifying reference to be primitive. However, it may be useful to introduce an interpretation concerning the use of 'identifying reference'.

In what follows, 'R' is used to express a ternary relation between a person, whose proper name may be assumed as a value by a variable

'α' ranging over a set of proper names for persons, and a space–time coordinate which is a value of a variable 'σ' taking as its values specific space–time coordinates. When identifying reference (hereafter simply called 'reference') to an object obtains, Rmo_is uniquely determines o_i in relation to a person m, at a certain space–time coordinate s:

$$(1) \qquad (x)(Rmxs \;\&\cdot x \in \{o_1, o_2, \ldots, o_n\} :\supset$$
$$-(\exists y)(Rmys \;\&\cdot y \in \{o_1, o_2, \ldots, o_n\} :\& \; x \neq y).$$[17]

From this point of view, the concept of reference is used to address the metalogical properties of identification; that is to say, possession of an identity is presupposed in connection with any particular, and all particulars are possible objects of reference, i.e., can be identified.[18] (It is important, then, to observe that the term 'reference' is *not* used in a way that entails the existence of psychological processes, intentions, etc., although these dimensions of referring need not be excluded if we wish to talk about them.)

Let $p_i \supset R\alpha o_i\sigma$ express the claim that the use by a person α at a space–time position σ of a referring sentence p_i entails reference to an o_i, if p_i has a value in the significant range; in other words, $R\alpha o_i\alpha$ follows from p_i whether the value of p_i is T or F.[19] The claim that is implicit here is that referring sentences of $\mathscr{P}$ are such that reference obtains to some o_i provided only that the p_i of $\mathscr{P}$ have truth-values in the significant range: hence, even when a p_i = F, reference is considered to obtain to some o_i which can serve to justify the claim to the effect that p_i = F.

A p_i is said to be *self-referentially inconsistent* in three cases which we distinguish here. (1) When $p_i \supset R\alpha o_i\sigma$ and $o_i = p_i$, then p_i exhibits *sentential* or *propositional self-reference*, depending upon whether p_i is considered as a sentence or as the expression of a proposition. If p_i is self-referential in either of these two ways and p_i claims of itself that it is false, then, when V is bivalent, p_i is true iff it is false. Such a p_i comprises a *paradox-generating* self-referential inconsistency. Many of the semantical paradoxes are clearly of this form.

(2) When $p_i \supset R\alpha o_i\sigma$ and $o_i = P_{p_i}$, where 'P' designates a pragmatic (or performatory) aspect of the use made of p_i by α at space–time position σ, then p_i is termed *pragmatically* (or performatively) *self-referential*. If p_i is pragmatically self-referential and p_i is such that if p_i is asserted or otherwise is used in a manner such that P_{p_i} falsifies

p_i, then, when V is bivalent, p_i is said to be *self-refuting*. The assertion, for example, "This assertion does not refer to an x such that Fx", for interpretations of 'x' and 'F', expresses a self-refuting self-referential inconsistency. Ramsey's familiar example, "I can't say 'cake'", when uttered by anyone, accordingly may be seen to be self-refuting.

(3) When $p_i \supset R\alpha o_i\sigma$ and $R\alpha o_i\sigma \supset R\alpha M_{p_i}\sigma$,[20] where '$M_{p_i}$' designates a "precondition of reference" which must be satisfied in order for p_i to have a value in the significant range, then p_i is termed *metalogically self-referential*. If p_i is metalogically self-referential and p_i is such that p_i denies one or more conditions which must be satisfied in order for it to be possible to assert, or otherwise use, p_i significantly, then p_i is said to be *projective*, or $\vec{p}_i$.[21]

The expression 'precondition of reference' is associated with the following equivalent senses: 'M_{p_i}' designates a "precondition of reference" if, in order for reference to be possible in a particular context of reference, M_{p_i} must be satisfied; M_{p_i} is a necessary condition of possible reference; M_{p_i} qualifies as a "precondition of reference" iff it designates a condition the non-satisfaction of which in a particular context of reference results in projection.

When V is bivalent, a metalogically self-referentially inconsistent p_i makes, with a putative value T or F, an ascription a of some object of reference o_i. If $p_i = \mathrm{T}$, then a applies to o_i, or $a(o_i)$; if $p_i = \mathrm{F}$, then $-a(o_i)$. In either case, possible reference to o_i is presupposed.

(2) $\qquad a(o_i) \vee -a(o_i). \supset \Diamond R\alpha o_i\sigma$

In short, when V is bivalent,

(3) $\qquad \vec{p}_i \equiv a(o_i) \vee -a(o_i). \;\&\; -\Diamond R\alpha o_i\sigma,$

where $-\Diamond R\alpha o_i\sigma$ is implied by the projective denial of one of the conditions which must be satisfied in order for it to be possible significantly to assert p_i.

$$\vdash p_{i(\mathrm{T}\vee\mathrm{F})},\ p_i \supset R\alpha o_i\sigma,\ R\alpha o_i\sigma \supset \Diamond R\alpha o_i\sigma,\ \Diamond R\alpha o_i\sigma \supset M_{p_i},$$

(4) $\qquad p_i \supset -M_{p_i} \vdash -\Diamond R\alpha o_i\sigma$

The self-referential inconsistency of a projection is rendered explicit when the consequent of (2) and the conclusion of (4) are conjoined.

P. W. Bridgman's hypothesis to the effect that the entire physical universe is shrinking homogeneously, i.e., in a manner such that all

operations of measurement are correspondingly affected, may be seen to be projective. In order for the hypothesis to be significant in a bivalent system, in order for reference to be made to "universal homogeneous shrinkage", Bridgman argues that it must be presupposed possible to detect relevant changes in relative size of the physical universe. This is essential to the meaning of the concept of shrinkage. However, by hypothesis, universal homogeneous shrinkage rules out that the precondition of reference, possible detection of the alleged change in relative size, can be satisfied. Hence the hypothesis is projective.[22]

In an intuitive sense, $p_i \supset \vec{p}_i$ will hold when p_i conflicts self-referentially with preconditions which must be granted in order for the value of p_i to fall in the significant range. A projective assertion consequently involves a special kind of self-referential inconsistency. Our main interest here is in projective forms of reference.

For a bivalent range of significance, $p_i = \mathrm{T}$ when

$$\text{`}p_i\text{'} \text{ is true iff } p_i \quad \text{(Tarski's definition);}$$

and $p_i = \mathrm{F}$ when

$$\text{`}p_i\text{'} \text{ is false iff } -p_i.$$

When p_i is projective, p_i is said to have value μ

$$\text{`}p_i\text{'} \text{ has value } \mu \text{ iff } \vec{p}_i.$$

Here, 'μ' represents the value 'projective meaninglessness' which lies outside the significant range of values $\{0, 1, \ldots, n-1\}$. It should be clear from the nature of a projective assertion that its value cannot be identified with any of the values in its putative significant range since one or more conditions are denied which must be satisfied in order for p_i to have *any* value in the significant range. The self-referential inconsistency of a projective assertion is of a kind which literally and logically precludes that the assertion *can* possess a value in the significant range. In some contexts there may be some latitude of choice whether to consider an assertion to be meaningless or false: e.g., in the case of the infamous 'The present king of France is bald'. From the standpoint of metalogic of reference, however, no other option is available: The value of a projective assertion must fall outside the significant range, hence it is appropriate to identify its value μ with meaninglessness.

A p_i is said to be *self-validating* in the case where $-p_i$ is metalogically self-referentially inconsistent. Conversely, p_i is said to be metalogically self-referentially inconsistent in the case where $-p_i$ is self-validating. I.e.,

(5) $(x)(x \in \mathscr{P} \cdot \& Fx :\supset \cdot G - x)$ and

$(x)(x \in \mathscr{P} \cdot \& Gx :\supset \cdot F - x)$,

where F is the property '... is self-validating' and G is the property '... is self-referentially inconsistent'.

It follows that for any $p_i \supset \bar{p}_i$, and hence when p_i has value μ, the equivalent claims 'the value of p_i does not fall in the significant range', 'p_i is not significant', 'p_i is meaningless' self-validate since the denial of any one entails self-referential inconsistency. *For this reason, referential consistency, as a metalogical criterion of meaning, cannot not be accepted.* Referential consistency is, in other words, a self-validating criterion which must be satisfied in order for claims to be meaningful.

It may be noted that the significant range of the set V of possible values of a p_i has been left unspecified, although in general we have defined the significant range to coincide with $\{0, 1, \ldots, n-1\}$. Leaving the significant range unspecified in this way has the advantage of flexibility, since, in some contexts of reference, we may wish to be able to assign values representing indeterminacy, statistical probabilities, etc., to a p_i (for example, in quantum logics). Although no decision has been made, then, in favor of bivalence in V, the following metametalanguage formulation is implied by the principle of bivalence, without implying it:

(i) Every referring sentence of $\mathscr{P}$ either has a value in the significant range, or it does not.

Adoption of this metalogical version of the principle of bivalence entails that all metalogical statements assigning values from $\{0, 1, \ldots, n-1, n\}$ – (from the range of possible values from falsity (0) to a designated value (1 in a bivalent system) to μ) – to a p_i are themselves true or they are not. In fact, (i) entails

(ii) There exist in principle possible procedures which yield a yes or no determination for any metalogical value-assigning statement about members of $\mathscr{P}$.

It will be evident to the reader that the assertion of (i) conjoined with the rejection of (ii) constitutes a projective assertion. Consequently, we shall regard (ii) as entailed, in a self-validating manner, by (i).[23]

By way of illustration, let us assume V is bivalent; hence the significant range comprises values T (1) and F (0) with μ representing the value of projective assertions. In effect, then, the set of sentences or propositions $\mathscr{P}' = \{p_1, p_2, \ldots, p_n\}$[24] will be, for the purposes of assessing referential consistency, three-valued within a bivalent metalanguage. (Such a three-valued representation can be reduced, as we shall see, to a two-valued representation, with T, F = Ψ, where 'Ψ' simply indicates a value in the significant range.)

Matrices for conjunction and negation suitably take the form proposed by Bochvar:[25]

$-$	
F	T
T	F
μ	μ

&	T	F	μ
T	T	F	μ
F	F	F	μ
μ	μ	μ	μ

Where μ is the value of a projective assertion, the above matrices make clear that the negation of a projection remains projective, while the conjunction of a projection with a significant assertion infects the compound statement, so to speak, with meaninglessness. The projective character of one conjunct may undermine the referential consistency of the other conjunct. The matrix for conjunction avoids this eventuality.

Other connectives are easily defined:

$A \vee B$ for $-(-A \,\&\, -B)$

$A \supset B$ for $-(A \,\&\, -B)$

$A \equiv B$ for $(A \supset B) \,\&\, (B \supset A)$,

so that the following matrices are determined:

$\vee$	T	F	μ
T	T	T	μ
F	T	F	μ
μ	μ	μ	μ

$\supset$	T	F	μ
T	T	F	μ
F	T	T	μ
μ	μ	μ	μ

$\equiv$	T	F	μ
T	T	F	μ
F	F	T	μ
μ	μ	μ	μ

From these matrices it can readily be seen that once part of an expression assumes the value μ, the expression automatically assumes the value μ. (The same rationale applies here as in the case of conjunction.) It is also evident that if all μ-rows and -columns are *eliminated*, the matrix is reduced to the normal two-valued one. If one sets $T, F = \Psi$, then it is clear that the elimination of statements of value μ leaves a set of statements having the value Ψ, statements which are in the significant range. This is an obviously desirable property of a necessary, not sufficient, criterion of meaning: its application will lead to the elimination of certain meaningless statements, leaving untouched all candidates which may be significant (and perhaps to which other necessary conditions of meaning may be applied.)

The metalogical criterion of meaning which emerges from this discussion is both *non-arbitrary* and *compelling*. It is *non-arbitrary* because the criterion is intrinsically informed by the special character of individual contexts of reference. It is *compelling* because one cannot at one and the same time consistently use expressions, sentences, or concepts referringly yet undermine their capacities to refer. Finally, a metalogical criterion of meaning which is defined in terms of referential consistency is *self-validating*: rejecting its application leads to projection.

In such a metalogical understanding of meaning, criteria for evaluating consistency and significance are determined as a function of one's needs and interests in making reference to certain kinds of objects. Within any specific context of reference, with these needs and interests in view, intrinsically determined criteria for evaluating internal consistency and significance *merge*, from the standpoint of a general metalogic of reference. They provide critical tools for appraising the meaningful use of expressions, sentences, or concepts in that context. Referential consistency is, in short, a contextually determined, yet non-arbitrary, compelling, and self-validating criterion of meaning.

In conclusion, it may be of interest to consider the relationship between a metalogical conception of meaning as a function of referential consistency, and the problem of putative meaningfulness.

THE PROBLEM OF PUTATIVE MEANINGFULNESS

Let p_i be a sentence or proposition in the context of a system SI which permits unambiguous identifying and re-identifying reference[26] to a set $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ of objects. Let it further be agreed that a p_i is asserted to have a truth-value in the significant range, i.e., $\neq \mu$.

Upon analysis, it is determined that $p_i \supset \vec{p}_i$, because $p_i \supset R\alpha o_i \sigma$, while $M_{pi} \supset -\Diamond R\alpha o_i \sigma$. From a metalogical frame of reference, M, then, we associate with p_i a truth-value μ not in the significant range. Note that *this* claim is an assertion about p_i-in-SI, and hence is a metalogical claim whose truth-value is determined on a bivalent metalogical basis.

It will be evident that the problem of putative meaningfulness is resolved. This problem has been pointed to by opponents to the use of meaning criteria. They have argued that, on the one hand, we have an expression, sentence, or concept which is used in various contexts, and in what is considered to be a meaningful fashion. Yet, upon application of a criterion of meaning, the alleged meaning is supposed to be given up, and the matter closed. The initially perceived meaning is not according to this view, "really meaningful". Such a suggestion runs counter to belief, i.e., is literally (not logically) paradoxical. This counter-intuitive character of results that stem from the application of meaning criteria decidedly has not promoted the popularity of criteria of meaning.

However, the quasi-paradoxical appearance of the problem of putative meaningfulness is simple to dispel: From the standpoint of SI, p_i is used to refer to an o_i so that o_i is uniquely determined. From the standpoint of M, reference is made to p_i-in-SI and reveals, through an analysis of p_i's referential preconditions, that the assertion of p_i-in-SI undermines p_i's capacity to refer to o_i.

If we associate a "meaning spectrum" V' with p_i such that $V' = \{0, 1, \ldots, n\}$, where $n = \mu$, then for any $0 \leq v_i < n$, v_i falls in the significant range V of V'. While the assignment of any v_i up to and including v_{n-1} may be made from the standpoint of SI, μ-assignments require recourse to a metalogical frame of reference M. In short, the possibility of detecting that a p_i has value μ is essentially a function of M's referential capacity. A metalogical statement S asserting that p_i is projective in SI, independently of M, itself is projective, as the reader may confirm.

There is, then, no problem with respect to putative meaningfulness from this viewpoint. What opponents to the use of meaning criteria very likely have in mind falls appropriately in this view under the heading of "making mistakes" and "detecting errors". When one makes a mistake without realizing it at the time, and later discovers his error, the passage of time provides what is, in effect, a metasystem which permits reference to what is retained in memory: From this vantage point, one compares what one remembers having thought earlier with what one now knows, and claims, in retrospect, that a mistake was made at the earlier time. The same may be said in the present case: The use of p_i to refer to o_i in SI was erroneous because p_i can be shown to be projective in M.[26] Hence, making an assertion which can be shown to be projective and hence meaningless in the sense developed, is simply to make one of many different kinds of possible mistakes.[27]

To remind us of this, it is convenient to view μ-assignments as involving, in a very literal sense, a *shift of significance.* Assumption of a metalogical frame of reference with respect to a projective assertion p_i results in a shift in p_i's putative truth-value (in SI) to μ (in M). Such a shift in significance is essentially a function of the metalogical frame of reference used. It is clear that a more comprehensive account of results proceeding from applications of a metalogic of reference would reveal many such *shifts to the value* μ of expressions, sentences, and concepts erroneously believed to be significant.[28]

Saint Louis University

NOTES

* Research reported here was supported in part by a grant from the Max-Planck-Gesellschaft.

[1] The general concern, to identify and eliminate meaningless statements and concepts from technical and/or ordinary discourse reflects a long tradition in which logic and philosophy together have sought to clarify our conceptual structure, and exhibit departures from sense. For example, Kant made mention of the need for a "negative science", a *phaenomenologia generalis*, which would undertake what might now be construed as a kind of "meaning-sorting", to insure that only meaningful propositions remain as the subject for subsequent analysis. (In a letter to Lambert, dated September 2, 1770.) The list of names in this tradition could be expanded almost indefinitely.

[2] For a statement of the view that logical rules essentially comprise conventions we

agree upon, and for additional references, see, e.g., Haskell B. Curry, *Outlines of a Formalist Philosophy of Mathematics* (Amsterdam: North-Holland 1957). See below, Note 12.

[3] Wittgenstein has given considerable attention to the relationship between using rules and achieving practical ends. See, e.g., his *Remarks on the Foundations of Mathematics*, edited by G. H. von Wright, R. Rhees, and G. E. M. Anscombe, trans. by G. E. M. Anscombe (Oxford: Basil Blackwell 1956): I – 9, 20, 131, 162; V – 31ff; and passim.

[4] Hume, *Enquiry*, sec. XII, iii.

[5] Moritz Schlick, 'Positivism and Realism', in A. J. Ayer (ed.), *Logical Positivism* (New York: Free Press 1959), pp. 82–107.

[6] A. J. Ayer, *Language, Truth and Logic* (London: Gollancz 1936; rev. ed. 1946), p. 35.

[7] Rudolf Carnap, *Philosophy and Logical Syntax* (London: Kegan Paul 1935), pp. 13–14.

[8] Carnap, *The Logical Structure of the World and Pseudoproblems in Philosophy*, trans. by R. George (Berkeley: University of California Press 1967), pp. 325ff and passim.

[9] There is certainly this sociological difference: Certain logical rules are "hard-programmed" in our culture, so that their *rejection* is counter-intuitive, as, for example, in the proposed use of non-distributive lattices in quantum theory. The matter is the other way around when it comes to criteria of meaning, since *violations* of the criteria heavily populate the domains of ordinary, and of some technical, discourse. And, to this extent, *acceptance* of certain of the proposed criteria of meaning frequently results in a counter-intuitive reaction in our culture.

[10] Carnap's introductory sentences in his *Logical Structure of the World* come to mind: "What is the purpose of a scientific book? It is meant to convince the reader of the validity of the thoughts which it presents."

[11] On the requirement that a theory of meaning be self-referentially meaningful, see R. J. Richman, 'On the Self-Reference of a Meaning Theory', *Philosophical Studies* **4** (1953), 69–72, and Paul F. Schmidt, 'Self-Referential Justification', *Philosophical Studies* **8** (1957), 49–54.

[12] I hasten to say, so as not to be misunderstood, that I am not principally concerned in this paper to recommend the formalists's thesis regarding the conventional nature of logical rules. However, viewing such rules in this way serves to highlight the contrast between them and the non-arbitrary and compelling criterion proposed here.

[13] See the author's 'The Idea of a Metalogic of Reference', *Methodology and Science: Interdisciplinary Journal for the Empirical Study of the Foundations of Science and their Methodology* **9** (1976), 85–92. Cf. also 'Phenomenology of the Implicit', *Dialectica* **29** (1975), 174–188, in which, from a phenomenological point of view, referential consistency permits both the identification of "projections" (see below in the text) and their elimination by means of a method of "de-projection", in a logically compelling manner. (For Polish readers, see 'Fenomenologia Tego, Co Implikowane', *Roczniki Filozoficzne* **22** (1974), 73–89.)

[14] A book is now in preparation, supported in part by the Max-Planck-Gesellschaft.

[15] For individual analyses which make use of referential consistency as a criterion of meaning, cf. the author's 'A Metatheoretical Basis for Interpretations of Problem Solving Behavior', *Methodology and Science* **11** (1978), 59–85, specifically §§10, 12;

'Towards a Unified Concept of Reality', *ETC.: A Review of General Semantics* **32** (1975), 43–49; 'Self-Reference, Phenomenology, and Philosophy of Science', *Methodology and Science* **13** (1980), 143–167.

A group of analyses in terms of referential consistency is detailed in *A Relativistic Theory of Phenomenological Constitution: A Self-Referential, Transcendental Approach to Conceptual Pathology* (English and French, 2 vols., Université de Paris 1970 – *Diss. Abs. Internatl.*, No. 79–05.)

[16] The convention is followed whereby False = 0, and the designated truth-value is $n - 1$; the value n is reserved for a purpose described later.

For generality, p_is with *variable* truth-value may be included: e.g., p_is for which value assignments are a function of time, as may be the case for future contingent statements, "So-and-so is alive", etc.

[17] It follows from this formulation that a person can refer identifyingly to only one object, of a set of possible objects of reference, at a time. The object referred to may be single or it may be compound, as when reference is made to a set having more than one member, or to a set of sets of objects, etc.

From the perspective presented here, when reference to an object o_i is uniquely determined, o_i is unambiguously identified in the sense of (1) in the text. The *identity* of o_i will essentially be a function of o_i's identifiability – hence, ultimately of frameworks relative to which reference to o_i can obtain.

A good deal must be omitted in this brief treatment: The possibility of re-identification would, for example, as Strawson has pointed out, need also to be assured.

[18] This recalls Quine's dictum, "no entity without identity." (Cf. Leonard Linsky, *Referring* (New York: Humanities Press 1967), p. 27.

[19] On the nature of '$\supset$' in such expressions as '$p_i \supset R\alpha o_i\sigma$', see Note 23 below.

[20] I.e., reference is made by α at σ to the (compound) object of reference $\{o_i, Mp_i\}$.

[21] The expression 'metalogical self-referential inconsistency' need not be restricted to the case in which reference obtains to $\{o_i, M_{p_i}\}$ at a single space-time σ. If $R\alpha o_i\sigma$, and $R\alpha M_{p_i}\sigma'$, σ' is later than σ, and $\vec{p}_i$, then we have the case where α realizes in retrospect that a p_i endorsed by him is projective, i.e., that in endorsing p_i at σ he was metalogically self-referentially inconsistent. Analogously, we may have the case where $R\alpha o_i\sigma$, $R\beta M_{p_i}\sigma'$, and σ' is later than σ: i.e., one man's commitments can be the basis of another man's metalogical analysis.

It is sometimes important to make a similar distinction in connection with pragmatic self-referential inconsistencies. Statements are sometimes and even frequently made by some individuals who are not aware at the time, and may never become aware, of the pragmatic self-referential inconsistencies they involve.

[22] For more detailed illustrations of projective forms of reference, see Note 15.

[23] The reader may be interested in contrasting the variety of entailment in question in this paper with "virtual implication" described by Hintikka, in which '$p \supset q$' is "self-sustaining": See J. Hintikka, *Knowledge and Belief: An Introduction to the Logic of the Two Notions* (Ithaca, N.Y.: Cornell University Press 1962), pp. 32, 57, and passim.

[24] $\mathscr{P}'$ includes $\mathscr{P}$ as a subset; $\mathscr{P}'$ contains in addition to p_is which fall in the significant range, p_is which have the value μ.

[25] A three-valued logic, in which the third value is 'meaninglessness' or 'undefined', is

used by Bochvar to stand for the value of paradox-generating propositions. Although his three-valued system is without a theory of types, it is nevertheless consistent. See D. A. Bochvar, 'Ob odnom tréhznačnom isčislénii i égo priménénii k analizu paradoksov klassičéskogo rasširénnogo funkcional'nogo iscisléniá' [On a three-valued logical calculus and its application to the analysis of contradictions], *Matématičéskij sbornik* **4** (1939), 287–308; and D. A. Bochvar, 'K voprosu o néprotivoréčivosti odnogo tréhznačnogo isčisléniá' [On the consistency of a three-valued calculus], *Matématicéskij sbornik* **12** (1943), 353–369; as well as Alonzo Church, 'Review of D. A. Bochvar's "On a three-valued logical calculus and its application to the analysis of contradictions"', *Journal of Symbolic Logic* **4** (1939), 98–99; with a correction in *Journal of Symbolic Logic* **5** (1940), 119.

Patrick Suppes makes use of Bochvar's three-valued system (without, however, crediting Bochvar for his truth-matrices) in connection with a formal representation of operationally meaningless statements. Cf. P. Suppes, 'Measurement, Empirical Meaningfulness, and Three-valued Logic', in P. Suppes, *Studies in the Methodology and Foundations of Science: Selected Papers from 1951 to 1969* (Dordrecht-Holland: D. Reidel 1969), pp. 65–79. (Reprinted from C. W. Churchman and P. Ratoosh (eds.), *Measurement: Definitions and Theories* (New York: John Wiley 1959), pp. 129–143.)

Several other authors have proposed three-valued systems in which the third value is 'meaninglessness'. For example: Sören Halldén, *The Logic of Nonsense* (Uppsala: Universitets Årsskrift II, **9** (1949)); Moh Shaw-kwei, 'Logical Paradoxes for Many-valued Systems', *Journal of Symbolic Logic* **19** (1954), 37–40; Lennart Åquist, 'Reflections on the Logic of Nonsense', *Theoria* **28** (1962), Part I, 138–157. For various reasons, however, special properties of these proposed systems make them unsuitable in the present context.

It might be mentioned that some authors have felt that the matrix for negation given in the text precludes a satisfactory interpretation of three-valued logic. That A and $-A$ have the same value when A has the value 'meaninglessness' seems to them problematic. Andrzej Mostowski, for example, has remarked in this connection that he does not have "any hope that it will ever be possible to find a reasonable interpretation of the three-valued logic of Łukasiewicz [which has the same matrix for negation as in Bochvar's system] in terms of ordinary language." A. Mostowski, 'Review of Helen Rasiowa's "A dziedziny logiki matematycznej. II. Logiki wielwartościowe Łukasiewicza"'. [From the domain of mathematical logic. II. The many-valued logic of Łukasiewicz], which appeared in *Journal of Symbolic Logic* **15** (1950), 223. Rasiowa's original paper appeared in *Matematyka* **3** (1950), 4–11.

It is, of course, my belief that Mostowski's pessimism was ill-founded.

[26] For an indication of the rationale behind the condition requiring that re-identification be possible in SI, see the text below, where this Note-number is repeated.

[27] Nothing need be said in any detail about what one's "intentions" may have been in using p_i in this way, since referring to what one had in mind but sees was not realized in actual expression, is in practice to orient oneself with respect to p_i-in-SI in the manner already described.

[28] Cf. Notes 13 and 15.

PART IV

COMPUTATIONAL SELF-REFERENCE

First Order Theories of Individual Concepts and Propositions

J. McCarthy
Computer Science Department
Stanford University, USA

Abstract

We discuss first order theories in which *individual concepts* are admitted as mathematical objects along with the things that *reify* them. This allows very straightforward formalizations of knowledge, belief, wanting, and necessity in ordinary first order logic without modal operators. Applications are given in philosophy and in artificial intelligence. We do not treat general concepts, and we do not present any full axiomatizations but rather show how various facts can be expressed.

INTRODUCTION

"*... it seems that hardly anybody proposes to use different variables for propositions and for truth-values, or different variables for individuals and individual concepts.*" – (Carnap 1956, p. 113).

Admitting individual concepts as objects – with concept-valued constants, variables, functions and expressions – allows ordinary first order theories of necessity, knowledge, belief and wanting without modal operators or quotation marks and without the restrictions on substituting equals for equals that either device makes necessary.

In this paper we will show how various individual concepts and propositions can be expressed. We are not yet ready to present a full collection of axioms. Moreover, our purpose is not to explicate what concepts are, in a philosophical sense, but rather to develop a language of concepts for representing facts about knowledge, belief, etc. in the memory of a computer.

Frege (1892) discussed the need to distinguish direct and indirect use of words. According to one interpretation of Frege's ideas, the meaning of the phrase "*Mike's telephone number*" in the sentence "*Pat knows Mike's telephone number*" is the concept of Mike's telephone number, whereas its meaning in the sentence "*Pat dialled Mike's telephone number*" is the number itself. Thus if

we also have "*Mary's telephone number = Mike's telephone number*", then "*Pat dialled Mary's telephone number*" follows, but "*Pat knows Mary's telephone number*" does not.

It was further proposed that a phrase has a *sense* which is a *concept* and is its *meaning* in *oblique contexts* like knowing and wanting, and a *denotation* which is its *meaning* in *direct contexts* like dialling. *Denotations* are the basis of the semantics of first order logic and model theory and are well understood, but *sense* has given more trouble, and the modal treatment of oblique contexts avoids the idea. On the other hand, logicians such as Carnap (1947 and 1956), Church (1951) and Montague (1974) see a need for *concepts* and have proposed formalizations. All these formalizations involve modifying the logic used; ours doesn't modify the logic and is more powerful, because it includes mappings from objects to concepts. Robert Moore's forthcoming dissertation also uses concepts in first order logic.

The problem identified by Frege – of suitably limiting the application of the substitutivity of equals for equals – arises in artificial intelligence as well as in philosophy and linguistics for any system that must represent information about beliefs, knowledge, desires, or logical necessity – regardless of whether the representation is declarative or procedural (as in PLANNER and other AI formalisms).

Our approach involves treating concepts as one kind of object in an ordinary first order theory. We shall have one term that denotes Mike's telephone number and a different term denoting the concept of Mike's telephone number instead of having a single term whose denotation is the number and whose sense is a concept of it. The relations among concepts and between concepts and other entities are expressed by formulas of first order logic. Ordinary model theory can then be used to study what spaces of concepts satisfy various sets of axioms.

We treat primarily what Carnap calls *individual concepts* like *Mike's telephone number* or *Pegasus* and not general concepts like *telephone* or *unicorn*. Extension to general concepts seems feasible, but individual concepts provide enough food for thought for the present.

This is a preliminary paper in that we don't give a comprehensive set of axioms for concepts. Instead we merely translate some English sentences into our formalism to give an idea of the possibilities.

KNOWING WHAT AND KNOWING THAT

To assert that Pat knows Mike's telephone number we write

$$true\ Know(Pat,\ Telephone\ Mike) \tag{1}$$

with the following conventions:

1. Parentheses are often omitted for one argument functions and predicates. This purely syntactic convention is not important. Another convention is to capitalize the first letter of a constant, variable, or function name

when its value is a concept. (We considered also capitalizing the last letter when the arguments are concepts, but it made the formulas ugly.)

2. *Mike* is the concept of Mike; that is, it is the *sense* of the expression "*Mike*". *mike* is Mike himself.
3. *Telephone* is a function that takes a concept of a person into a concept of his telephone number. We will also use *telephone* which takes the person himself into the telephone number itself. We do not propose to identify the function *Telephone* with the general concept of a person's telephone number.
4. If P is a person concept and X is another concept, then *Know*(P, X) is an assertion concept or *proposition* meaning that P *knows* the value of X. Thus in (1) *Know*(*Pat*, *Telephone Mike*) is a proposition and not a truth value. Note that we are formalizing *knowing what* rather than *knowing that* or *knowing how*. For AI and for other practical purposes, *knowing what* seems to be the most useful notion of the three. In English, *knowing what* is written *knowing whether* when the "knowand" is a proposition.
5. It is often convenient to write *know*(*pat*, *Telephone Mike*) instead of *true Know*(*Pat*, *Telephone Mike*) when we don't intend to iterate knowledge further. *know* is a predicate in the logic, so we cannot apply any knowledge operators to it. We will have

$$know(pat, Telephone\ Mike) \equiv true\ Know(Pat, Telephone\ Mike). \qquad (2)$$

6. We expect that the proposition *Know*(*Pat*, *Telephone Mike*) will be useful accompanied by axioms that allow inferring that Pat will use this knowledge under appropriate circumstances, that is, he will dial it or retell it when appropriate. There will also be axioms asserting that he will know it after being told it or looking it up in the telephone book.
7. While the sentence "*Pat knows Mike*" is in common use, it is harder to see how *Know*(*Pat*, *Mike*) is to be used and axiomatized. I suspect that new methods will be required to treat knowing a person.
8. *true* Q is the truth value, t or f, of the proposition Q, and we must write *true* Q in order to assert Q. Later we will consider formalisms in which *true* has a another argument – a *situation*, a *story*, a *possible world*, or even a *partial possible world* (a notion we suspect will eventually be found necessary).
9. The formulas are in a sorted first order logic with functions and equality. Knowledge, necessity, etc. will be discussed without extending the logic in any way – solely by the introduction of predicate and function symbols subject to suitable axioms. In the present informal treatment, we will not be explicit about sorts, but we will use different letters for variables of different sorts.

The reader may be nervous about what is meant by *concept*. He will have to remain nervous; no final commitment will be made in this paper. The formalism is compatible with many possibilities, and these can be compared by using the models of their first order theories. Actually, this paper isn't much motivated by the philosophical question of what concepts really are. The goal is more to make a formal structure that can be used to represent facts about knowledge and belief so that a computer program can reason about who has what knowledge in order to solve problems. From either the philosophical or the AI point of view, however, if (1) is to be reasonable, it must not follow from (1) and the fact that Mary's telephone number is the same as Mike's, that Pat knows Mary's telephone number.

The proposition that Joe knows *whether* Pat knows Mike's telephone number, is written

$$Know(Joe, Know(Pat, Telephone\ Mike)), \tag{3}$$

and asserting it requires writing

$$true\ Know(Joe, Know(Pat, Telephone\ Mike)), \tag{4}$$

while the proposition that Joe knows *that* Pat knows Mike's telephone number is written

$$K(Joe, Know(Pat, Telephone\ Mike)), \tag{5}$$

where $K(P, Q)$ is the proposition that P knows *that* Q. English does not treat knowing a proposition and knowing an individual concept uniformly: knowing an individual concept means knowing its value, while knowing a proposition means knowing that it has a particular value, namely t. There is no reason to impose this infirmity on robots.

We first consider systems in which corresponding to each concept X, there is a thing x of which X is a concept. Then there is a function *denot* such that

$$x = denot\ X. \tag{6}$$

Functions like *Telephone* are then related to *denot* by equations like

$$\forall P1\ P2.(denot\ P1 = denot\ P2 \supset denot\ Telephone\ P1 = denot\ Telephone\ P2). \tag{7}$$

We call *denot X* the *denotation* of the concept X, and (7) asserts that the denotation of the concept of P's telephone number depends only on the denotation of the concept P. The variables in (7) range over concepts of persons, and we regard (7) as asserting that *Telephone* is *extensional* with respect to *denot*. Note that our *denot* operates on concepts rather than on expressions; a theory of expressions will also need a denotation function. From (7) and suitable logical axioms follows the existence of a function *telephone* satisfying

$$\forall P.(denot\ Telephone\ P = telephone\ denot\ P). \tag{8}$$

Know is extensional with respect to *denot* in its first argument, and this is expressed by

$$\forall P1\ P2\ X.(denot\ P1 = denot\ P2 \supset denot\ Know(P1, X) = denot\ Know(P2, X)), \tag{9}$$

but it is not extensional in its second argument. We can therefore define a predicate $know(p, X)$ satisfying

$$\forall P\ X.(true\ Know(P, X) \equiv know(denot\ P, X)). \tag{10}$$

(Note that all these predicates and functions are entirely extensional in the underlying logic, and the notion of extensionality presented here is relative to *denot*.)

The predicate *true* and the function *denot* are related by

$$\forall Q.(true\ Q \equiv (denot\ Q = t)) \tag{11}$$

provided that truth values are in the range of *denot*, and *denot* could also be provided with a (*partial*) *possible world* argument.

When we don't assume that all concepts have denotations, we use a predicate $denotes(X, x)$ instead of a function. The extensionality of *Telephone* would then be written

$$\forall P1\ P2\ x\ u.(denotes(P1, x) \wedge denotes(P2, x) \wedge denotes(Telephone\ P1, u) \supset denotes(Telephone\ P2, u)). \tag{12}$$

We now introduce the function *Exists* satisfying

$$\forall X.(true\ Exists\ X \equiv \exists x.denotes(X, x)). \tag{13}$$

Suppose we want to assert that Pegasus is a horse without asserting that Pegasus exists. We can do this by introducing the predicate *Ishorse* and writing

$$true\ Ishorse\ Pegasus \tag{14}$$

which is related to the predicate *ishorse* by

$$\forall X\ x.(denotes(X, x) \supset (ishorse\ x \equiv true\ Ishorse\ X)). \tag{15}$$

In this way, we assert extensionality without assuming that all concepts have denotations. *Exists* is extensional in this sense, but the corresponding predicate *exists* is identically true and therefore dispensable.

To combine concepts propositionally, we need analogs of the propositional operators such as *And*, which we shall write as an infix and axiomatize by

$$\forall Q1\ Q2.(true(Q1\ And\ Q2) \equiv true\ Q1 \wedge true\ Q2). \tag{16}$$

The corresponding formulas for *Or*, *Not*, *Implies*, and *Equiv* are

$$\forall Q1\ Q2.(true(Q1\ Or\ Q2) \equiv true\ Q1 \vee true\ Q2), \tag{17}$$

$$\forall Q.(true(Not\ Q) \equiv \neg\ true\ Q), \tag{18}$$

$$\forall Q1\ Q2.(true(Q1\ Implies\ Q2) \equiv true\ Q1 \supset true\ Q2), \tag{19}$$

and

$$\forall Q1\ Q2.(true(Q1\ Equiv\ Q2) \equiv (true\ Q1 \equiv true\ Q2)). \tag{20}$$

The equality symbol "=" is part of the logic so that $X = Y$ asserts that X and Y are the same concept. To write propositions expressing equality, we introduce $Equal(X, Y)$ which is a proposition that X and Y denote the same thing if anything. We shall want axioms

$$\forall X.\ true\ Equal(X, X), \tag{21}$$

$$\forall X\ Y.(true\ Equal(X, Y) \equiv true\ Equal(Y, X)), \tag{22}$$

and

$$\forall X\ Y\ Z.(true\ Equal(X, Y) \wedge true\ Equal(Y, Z) \supset true\ Equal(X, Z) \tag{23}$$

making $true\ Equal(X, Y)$ an equivalence relation, and

$$\forall X\ Y\ x.(true\ Equal(X, Y) \wedge denotes(X, x) \supset denotes(Y, x)) \tag{24}$$

which relates it to equality in the logic. We can make the concept of equality *essentially* symmetric by replacing (22) by

$$\forall X\ Y.\ Equal(X, Y) = Equal(Y, X), \tag{25}$$

that is, making the two expressions denote the *same concept.*

The statement that Mary has the same telephone number as Mike is asserted by

$$true\ Equal(Telephone\ Mary, Telephone\ Mike), \tag{26}$$

and it obviously doesn't follow from this and (1) that

$$true\ Know(Pat, Telephone\ Mary). \tag{27}$$

To draw this conclusion we need something like

$$true\ K(Pat, Equal(Telephone\ Mary, Telephone\ Mike)) \tag{28}$$

and suitable axioms about knowledge.

If we were to adopt the convention that a proposition appearing at the outer level of a sentence is asserted and were to regard the denotation-valued function as standing for the sense-valued function when it appears as the second argument of *Know*, we would have a notation that resembles ordinary language in handling obliquity entirely by context. There is no guarantee that general statements could be expressed unambiguously without circumlocution; the fact that the principles of intensional reasoning haven't yet been stated is evidence against the suitability of ordinary language for stating them.

FUNCTIONS FROM THINGS TO CONCEPTS OF THEM

While the relation *denotes*(X, x) between concepts and things is many-one, functions going from things to certain concepts of them seem useful. Some things such as numbers can be regarded as having *standard* concepts. Suppose that *Concept*1 n gives a standard concept of the number n, so that

$$\forall n.(denot\ Concept1\ n = n). \tag{29}$$

We can then have simultaneously

$$true\ Not\ Knew(Kepler,\ Composite\ Number\ Planets) \tag{30}$$

and

$$true\ Knew(Kepler,\ Composite\ Concept1\ denot\ Number\ Planets). \tag{31}$$

(We have used *Knew* instead of *Know*, because we are not now concerned with formalizing tense.)

(31) can be condensed using *Composite*1 which takes a number into the proposition that it is composite, that is,

$$Composite1\ n \stackrel{\wedge}{=} Composite\ Concept1\ n \tag{32}$$

getting

$$true\ Knew(Kepler,\ Composite1\ denot\ Number\ Planets). \tag{33}$$

A further condensation can be achieved by using *Composite*2 defined by

$$Composite2\ N = Composite\ Concept1\ denot\ N, \tag{34}$$

letting us write

$$true\ Knew(Kepler,\ Composite2\ Number\ Planets), \tag{35}$$

which is true even though

$$true\ Knew(Kepler,\ Composite\ Number\ Planets) \tag{36}$$

is false. (36) is our formal expression of "*Kepler knew that the number of planets is composite*", while (31), (33), and (35) each expresses a proposition that can only be stated awkwardly in English, for example, as "*Kepler knew that a certain number is composite, where this number (perhaps unbeknownst to Kepler) is the number of planets*".

We may also want a map from things to concepts of them in order to formalize a sentence like, "*Lassie knows the location of all her puppies*". We write this

$$\forall x.(ispuppy(x, lassie) \supset true\ Knowd(Lassie,\ Locationd\ Conceptd\ x)). \tag{37}$$

Here *Conceptd* takes a puppy into a dog's concept of it, and *Locationd* takes a dog's concept of a puppy into a dog's concept of its location. The axioms

satisfied by *Knowd*, *Locationd* and *Conceptd* can be tailored to our ideas of what dogs know.

A suitable collection of functions from things to concepts might permit a language that omitted some individual concepts like *Mike* (replacing it with *Conceptx mike*) and wrote many sentences with quantifiers over things rather than over concepts. However, it is still premature to apply Occam's razor. It may be possible to avoid concepts as objects in expressing particular facts but impossible to avoid them in stating general principles.

RELATIONS BETWEEN KNOWING WHAT AND KNOWING THAT

As mentioned before, "*Pat knows Mike's telephone number*" is written

$$true\ Know(Pat, Telephone\ Mike). \tag{38}$$

We can write "*Pat knows Mike's telephone number is* 333-3333"

$$true\ K(Pat, Equal(Telephone\ Mike, Concept1\ \text{"333-3333"})) \tag{39}$$

where $K(P, Q)$ is the proposition that $denot(P)$ knows the proposition Q and $Concept1$ ("333-3333") is some standard concept of that telephone number.

The two ways of expressing knowledge are somewhat interdefinable, since we can write

$$K(P, Q) = (Q\ And\ Know(P, Q)), \tag{40}$$

and

$$true\ Know(P, X) \equiv \exists A.(constant\ A \wedge true\ K(P, Equal(X, A))). \tag{41}$$

Here *constant A* asserts that A is a constant, that is, a concept such that we are willing to say that P knows X if he knows it equals A. This is clear enough for some domains like integers, but it is not obvious how to treat knowing a person.

Using the *standard concept* function $Concept1$, we might replace (41) by

$$true\ Know(P, X) \equiv \exists a.true\ K(P, Equal(X, Concept1\ a)) \tag{42}$$

with similar meaning.

(41) and (42) expresses a *denotational* definition of *Know* in terms of K. A *conceptual* definition seems to require something like

$$\forall P\ X.(Know(P, X) = Exists\ X\ And\ K(P, Equal(X, Concept2\ denot\ X))), \tag{43}$$

where $Concept2$ is a suitable function from things to concepts and may not be available for all sorts of objects.

REPLACING MODAL OPERATORS BY MODAL FUNCTIONS

Using concepts we can translate the content of modal logic into ordinary logic. We need only replace the modal operators by *modal functions*. The axioms of modal logic then translate into ordinary first order axioms. In this section we

will treat only *unquantified modal logic*. The arguments of the modal functions will not involve quantification, although quantification occurs in the outer logic.

Nec Q is the proposition that the proposition *Q* is necessary, and *Poss Q* is the proposition that it is possible. To assert necessity or possibility we must write *true Nec Q* or *true Poss Q*. This can be abbreviated by defining *nec Q* ≡ *true Nec Q* and *poss Q* correspondingly. However, since *nec* is a predicate in the logic with *t* and *f* as values, *nec Q* cannot be an argument of *nec* or *Nec*.

Before we even get to modal logic proper we have a decision to make – shall *Not Not Q* be considered the same proposition as *Q*, or is it merely extensionally equivalent? The first is written

$$\forall Q.\, Not\ Not\ Q = Q. \tag{44}$$

and the second

$$\forall Q.\, true\ Not\ Not\ Q \equiv true\ Q. \tag{45}$$

The second follows from the first by substitution of equals for equals.

In *Meaning and Necessity*, Carnap takes what amounts to the first alternative, regarding concepts as L-equivalence classes of expressions. This works nicely for discussing necessity, but when he wants to discuss knowledge without assuming that everyone knows Fermat's last theorem if it is true, he introduces the notion of *intensional isomorphism* and has knowledge operate on the equivalence classes of this relation.

If we choose the first alternative, then we may go on to identify any two propositions that can be transformed into each other by Boolean identities. This can be assured by a small collection of propositional identities like (44) including associative and distributive laws for conjunction and disjunction, De Morgan's law, and the laws governing the propositions *T* and *F*. In the second alternative we will want the extensional forms of the same laws. When we get to quantification a similar choice will arise, but if we choose the first alternative, it will be undecidable whether two expressions denote the same concept. I doubt that considerations of linguistic usage or usefulness in AI will unequivocally recommend one alternative, so both will have to be studied.

Actually there are more than two alternatives. Let *M* be the free algebra built up from the "atomic" concepts by the concept forming function symbols. If ≡≡ is an equivalence relation on *M* such that

$$\forall X1\ X2 \in M.((X1 \equiv\equiv X2) \supset (true\ X1 \equiv true\ X2)), \tag{46}$$

then the set of equivalence classes under ≡≡ may be taken as the set of concepts.

Similar possibilities arise in modal logic. We can choose between the *conceptual identity*

$$\forall Q.(Poss\ Q = Not\ Nec\ Not\ Q), \tag{47}$$

and the weaker extensional axiom

$$\forall Q.(true\ Poss\ Q \equiv true\ Not\ Nec\ Not\ Q). \tag{48}$$

We will write the rest of our modal axioms in extensional form.

We have

$$\forall Q.(true\ Nec\ Q \supset true\ Q), \tag{49}$$

and

$$\forall Q1\ Q2. \quad (true\ Nec\ Q1 \wedge true\ Nec(Q1\ Implies\ Q2) \supset true\ Nec\ Q2). \tag{50}$$

yielding a system equivalent to von Wright's T.

S4 is given by

$$\forall Q.(true\ Nec\ Q \equiv true\ Nec\ Nec\ Q), \tag{51}$$

and S5 by

$$\forall Q.(true\ Poss\ Q \equiv true\ Nec\ Poss\ Q). \tag{52}$$

Actually, there may be no need to commit ourselves to a particular modal system. We can simultaneously have the functions *NecT*, *Nec*4 and *Nec* 5, related by axioms such as

$$\forall Q.(true\ Nec4\ Q \supset true\ Nec5\ Q) \tag{53}$$

which would seem plausible if we regard S4 as corresponding to provability in some system and S5 as truth in the intended model of the system.

Presumably we shall want to relate necessity and equality by the axiom

$$\forall X.true\ Nec\ Equal(X, X). \tag{54}$$

Certain of Carnap's proposals translate to the stronger relation

$$\forall X\ Y.(X = Y \equiv true\ Nec\ Equal(X, Y)) \tag{55}$$

which asserts that two concepts are the same if and only if the equality of what they may denote is necessary.

MORE PHILOSOPHICAL EXAMPLES – MOSTLY WELL KNOWN

Some sentences that recur as examples in the philosophical literature will be expressed in our notation, so the treatments can be compared.

First we have "*The number of planets* = 9" and "*Necessarily* 9 = 9" from which one doesn't want to deduce "*Necessarily the number of planets* = 9". This example is discussed by Quine (1961) and (Kaplan 1969). Consider the sentences

$$\neg nec\ Equal(Number\ Planets, Concept1\ 9) \tag{56}$$

and

$$nec\ Equal(Concept1\ number\ planets, Concept1\ 9). \tag{57}$$

Both are true. (56) asserts that it is not necessary that the number of planets be 9, and (57) asserts that the number of planets, once determined, is the number

that is necessarily equal to 9. It is a major virtue of our formalism that both meanings can be expressed and are readily distinguished. Substitutivity of equals holds in the logic but causes no trouble, because "*The number of planets* = 9" may be written

$$number(planets) = 9 \tag{58}$$

or, using concepts as

$$true\ Equal(Number\ Planets, Concepts1\ 9), \tag{59}$$

and "*Necessarily* 9=9" is written

$$nec\ Equal(Concept1\ 9, Concept1\ 9), \tag{60}$$

and these don't yield the unwanted conclusion.

Ryle used the sentences "*Baldwin is a statesman*" and "*Pickwick is a fiction*" to illustrate that parallel sentence construction does not always give parallel sense. The first can be rendered in four ways, namely *true Statesman Baldwin* or *statesman denot Baldwin* or *statesman baldwin* or *statesman1 Baldwin* where the last asserts that the concept of Baldwin is one of a statesman. The second can be rendered only as *true Fiction Pickwick* or *fiction1 Pickwick*.

Quine (1961) considers illegitimate the sentence

$$(\exists x)(Philip\ is\ unaware\ that\ x\ denounced\ Catiline) \tag{61}$$

obtained from "*Philip is unaware that Tully denounced Catiline*" by existential generalization. In the example, we are also supposing the truth of *Philip is aware that Cicero denounced Cataline*". These sentences are related to (perhaps even explicated by) several sentences in our system. *Tully* and *Cicero* are taken as distinct concepts. The person is called *tully* or *cicero* in our language, and we have

$$tully = cicero, \tag{62}$$

$$denot\ Tully = cicero \tag{63}$$

and

$$denot\ Cicero = cicero. \tag{64}$$

We can discuss Philip's concept of the person Tully by introducing a function $Concept2(p1, p2)$ giving for some persons $p1$ and $p2$, $p1$'s concept of $p2$. Such a function need not be unique or always defined, but in the present case, some of our information may be conveniently expressed by

$$Concept2(philip, tully) = Cicero, \tag{65}$$

asserting that Philip's concept of the person Cicero is *Cicero*. The basic assumptions of Quine's example also include

$$true\ K(Philip, Denounced(Cicero, Catiline)) \tag{66}$$

and

$$\neg true\ K(Philip, Denounced(Tully, Catiline)). \tag{67}$$

From (63), ... (67) we can deduce

$$\exists P.\, true\ Denounced(P, Catiline)\ And\ Not\ K(Philip, Denounced(P, Catiline)), \tag{68}$$

from (67) and others, and

$$\neg\exists p.(denounced(p, catiline) \wedge \neg true\ K(Philip, Denounced(Concept2(philip, p), Catiline))) \tag{69}$$

using the additional hypotheses

$$\forall p.(denounced(p, catiline) \supset p = cicero), \tag{70}$$

$$denot\ Catiline = catiline, \tag{71}$$

and

$$\forall P1\ P2.(denot\ Denounced(P1, P2) \equiv denounced(denot\ P1, denot\ P2)). \tag{72}$$

Presumably (68) is always true, because we can always construct a concept whose denotation is Cicero unbeknownst to Philip. The truth of (69) depends on Philip's knowing that someone denounced Catiline, and on the map $Concept2(p1, p2)$ that gives one person's concept of another. If we refrain from using a silly map that gives something like $Denouncer(Catiline)$ as its value, we can get results that correspond to intuition.

The following sentence attributed to Russell is discussed by Kaplan: "*I thought that your yacht was longer than it is*". We can write it

$$true\ Believed(I, Greater(Length\ Youryacht, Concept1\ denot\ Length\ Youryacht)) \tag{73}$$

where we are not analysing the pronouns or the tense, but are using *denot* to get the actual length of the yacht and *Concept*1 to get back a concept of this true length so as to end up with a proposition that the length of the yacht is greater than that number. This looks problematical, but if it is consistent, it is probably useful.

To express "*Your yacht is longer than Peter thinks it is.*" we need the expression $Denot(Peter, X)$ giving a concept of what Peter thinks the value of X is. We now write

$$longer(youryacht, denot\ Denot(Peter, Length\ Youryacht)), \tag{74}$$

but I am not certain this is a correct translation.

Quine (1956) discusses an example in which Ralph sees Bernard J. Ortcutt skulking about and concludes that he is a spy, and also sees him on the beach, but doesn't recognize him as the same person. The facts can be expressed in our formalism by equations

$$true\ Believe(Ralph, Isspy\ P1) \tag{75}$$

and

$$true\ Believe(Ralph, Not\ Isspy\ P2) \tag{76}$$

where $P1$ and $P2$ are concepts satisfying *denot* $P1$ = *ortcutt* and *denot* $P2$ = *ortcutt*. $P1$ and $P2$ are further described by sentences relating them to the circumstances under which Ralph formed them.

We can still consider a simple sentence involving the persons as things -- write it *believespy*(*ralph*, *ortcutt*), where we define

$$\forall p1\, p2.(believespy(p1,p2) \equiv \\ true\ Believe(Concept1\ p1, Isspy\ Concept7\ p2) \qquad (77)$$

using suitable mappings *Concept*1 and *Concept*7 from persons to concepts of persons. We might also choose to define *believespy* in such a way that it requires *true Believe*(*Concept*1 $p1$, *Isspy* P) for several concepts P of $p2$, for example, the concepts arising from all $p1$'s encounters with $p2$ or his name. In this case *believespy*(*ralph*, *ortcutt*) will be false and so would a corresponding *notbelievespy*(*ralph*, *ortcutt*). However, the simple-minded predicate *believespy*, suitably defined, may be quite useful for expressing the facts necessary to predict someone's behaviour in simpler circumstances.

Regarded as an attempt to explicate the sentence "*Ralph believes Ortcutt is a spy*", the above may be considered rather tenuous. However, we are proposing it as a notation for expressing Ralph's beliefs about Ortcutt so that correct conclusions may be drawn about Ralph's future actions. For this it seems to be adequate.

PROPOSITIONS EXPRESSING QUANTIFICATION

As the examples of the previous sections have shown, admitting concepts as objects and introducing standard concept functions makes "quantifying in" rather easy. However, forming propositions and individual concepts by quantification requires new ideas and additional formalism. We are not very confident of the approach presented here.

We want to continue describing concepts within first order logic with no logical extentions. Therefore, in order to form new concepts by quantification and description, we introduce functions *All*, *Exist*, and *The* such that *All*(V, P) is (approximately) the proposition that *for all values of* V P *is true*, *Exist*(V, P) is the corresponding existential proposition, and *The*(V, P) is the concept of *the* V *such that* P.

Since *All* is to be a function, V and P must be objects in the logic. However, V is semantically a variable in the formation of *All*(V, P), *etc.*, and we will call such objects *inner variables* so as to distinguish them from variables in the logic. We will use V, sometimes with subscripts, for a logical variable ranging over inner variables. We also need some constant symbols for inner variables (got that?), and we will use doubled letters, sometimes with subscripts, for these. XX will be used for individual concepts, PP for persons, and QQ for propositions.

The second argument of *All* and friends is a "proposition with variables in it", but remember that these variables are inner variables which are constants

in the logic. Got that? We won't introduce a special term for them, but will generally allow concepts to include inner variables. Thus concepts now include inner variables like *XX* and *PP*, and concept-forming functions like *Telephone* and *Know* take the generalized concepts as arguments.

Thus

Child(*Mike*, *PP*) *Implies Equal*(*Telephone PP*, *Telephone Mike*) (78)

is a proposition with the inner variable *PP* in it to the effect that if *PP* is a child of Mike, then his telephone number is the same as Mike's, and

All(*PP*, *Child*(*Mike*, *PP*)
Implies Equal(*Telephone PP*, *Telephone Mike*)) (79)

is the proposition that all Mike's children have the same telephone number as Mike. Existential propositions are formed similarly to universal ones, but the function *Exist* introduced here should not be confused with the function *Exists* applied to individual concepts introduced earlier.

In forming individual concepts by the description function *The*, it doesn't matter whether the object described exists. Thus

The(*PP*, *Child*(*Mike*, *PP*)) (80)

is the concept of Mike's only child. *Exists The*(*PP*, *Child*(*Mike*, *PP*)) is the proposition that the described child exists. We have

true Exists The(*PP*, *Child*(*Mike*, *PP*)) ≡
true(*Exist*(*PP*, *Child*(*Mike*, *PP*)
And All(*PP*1, *Child*(*Mike*, *PP*1) *Implies Equal*(*PP*, *PP*1)))), (81)

but we may want the equality of the two propositions, that is,

Exists The(*PP*, *Child*(*Mike*, *PP*)) =
Exist(*PP*, *Child*(*Mike*, *PP*)
And All(*PP*1, *Child*(*Mike*, *PP*1) *Implies Equal*(*PP*, *PP*1))). (82)

This is part of general problem of when two logically equivalent concepts are to be regarded as the same.

In order to discuss the truth of propositions and the denotation of descriptions, we introduce *possible worlds* reluctantly and with an important difference from the usual treatment. We need them to give values to the inner variables, and we can also use them for axiomatizing the modal operators, knowledge, belief and tense. However, for axiomatizing quantification, we also need a function α such that

$$\pi' = \alpha(V, x, \pi) \tag{83}$$

is the possible world that is the same as the world π except that the inner variable V has the value x instead of the value it has in π. In this respect our possible worlds resemble the *state vectors* or *environments* of computer science

more than the possible worlds of the Kripke treatment of modal logic. This Cartesian product structure on the space of possible worlds can also be used to treat conterfactual conditional sentences.

Let $\pi 0$ be the actual world. Let $true(P, \pi)$ mean that the proposition P is true in the possible world π. Then

$$\forall P.(true\ P \equiv true(P, \pi 0)). \tag{84}$$

Let $denotes(X, x, \pi)$ mean that X denotes x in π, and let $denot(X, \pi)$ mean the denotation of X in π when that is defined.

The truth condition for $All(V, P)$ is then given by

$$\forall \pi\ V\ P.(true(All(V, P), \pi) \equiv \forall x.true(P, \alpha(V, x, \pi)). \tag{85}$$

Here V ranges over inner variables, P ranges over propositions, and x ranges over things. There seems to be no harm in making the domain of x depend on π. Similarly

$$\forall \pi\ V\ P.(true(Exist(V, P), \pi) \equiv \exists x.true(P, \alpha(V, x, \pi)). \tag{86}$$

The meaning of $The(V, P)$ is given by

$$\forall \pi\ V\ P\ x.(true(P, \alpha(V, x, \pi)) \wedge \forall y.(true(P, \alpha(V, y, \pi)) \supset y = x) \supset denotes(The(V, P), x, \pi)) \tag{87}$$

and

$$\forall \pi\ V\ P.(\neg \exists\ x.true(P, \alpha(V, x, \pi)) \supset \neg true\ Exists\ The(V, P)). \tag{88}$$

We also have the following "syntactic" rules governing propositions involving quantification:

$$\forall \pi\ Q1\ Q2\ V.(absent(V, Q1) \wedge true(All(V, Q1\ Implies\ Q2), \pi) \supset true(Q1\ Implies\ All(V, Q2), \pi)) \tag{89}$$

and

$$\forall \pi\ V\ Q\ X.(true(All(V, Q), \pi) \supset true(Subst(X, V, Q), \pi)). \tag{90}$$

where $absent(V, X)$ means that the variable V is not present in the concept X, and $Subst(X, V, Y)$ is the concept that results from substituting the concept X for the variable V in the concept Y. *absent* and *Subst* are characterized by the following axioms:

$$\forall V1\ V2.(absent(V1, V2) \equiv V1 \neq V2), \tag{91}$$

$$\forall V\ P\ X.(absent(V, Know(P, X)) \equiv absent(V, P) \wedge absent(V, X)), \tag{92}$$

axioms similar to (92) for other conceptual functions,

$$\forall V\ Q.absent(V, All(V, Q)), \tag{93}$$

$$\forall V\ Q.absent(V, Exist(V, Q)), \tag{94}$$

$$\forall V\ Q.absent(V, The(V, Q)), \tag{95}$$

$$\forall V\ X.Subst(V, V, X) = X, \tag{96}$$

$$\forall X\ V.Subst(X, V, V) |= X, \tag{97}$$

$$\forall X\ V\ P\ Y.(Subst(X, V, Know(P, Y)) = Know(Subst(X, V, P), Subst(X, V, Y))), \tag{98}$$

axioms similar to (98) for other functions,

$$\forall X\ V\ Q.(absent(V, Y) \supset Subst(X, V, Y) = Y), \tag{99}$$

$$\forall X\ V1\ V2\ Q.(V1 \neq V2 \wedge absent(V2, X) \supset Subst(X, V1, All(V2, Q)) = All(V2, Subst(X, V1, Q))), \tag{100}$$

and corresponding axioms to (100) for *Exist* and *The*.

Along with these comes the axiom that binding kills variables, that is,

$$\forall V1\ V2\ Q.(All(V1, Q) = All(V2, Subst(V2, V1, Q))). \tag{101}$$

The functions *absent* and *Subst* play a "syntactic" role in describing the rules of reasoning and don't appear in the concepts themselves. It seems likely that this is harmless until we want to form concepts of the laws of reasoning.

We used the Greek letter π for possible worlds, because we did not want to consider a possible world as a thing and introduce concepts of possible worlds. Reasoning about reasoning may require such concepts or else a formulation that doesn't use possible worlds.

Martin Davis (in conversation) pointed out the advantages of an alternative treatment avoiding possible worlds in case there is a single domain of individuals each of which has a standard concept. Then we can write

$$\forall V\ Q.(true\ All(V, Q) \equiv \forall x.true\ Subst(Concept1\ x, V, Q)). \tag{102}$$

POSSIBLE APPLICATIONS TO ARTIFICIAL INTELLIGENCE

The foregoing discussion of concepts has been mainly concerned with how to translate into a suitable formal language certain sentences of ordinary language. The sucess of the formalization is measured by the extent to which the logical consequences of these sentences in the formal system agree with our intuitions of what these consequences should be. Another goal of the formalization is to develop an idea of what concepts really are, but the possible formalizations have not been explored enough to draw even tentative conclusions about that.

For artificial intelligence, the study of concepts has yet a different motivation. Our success in making computer programs with *general intelligence* has been extremely limited, and one source of the limitation is our inability to formalize what the world is like in general. We can try to separate the problem of describing the general aspects of the world from the problem of using such a description and the facts of a situation to discover a strategy for achieving a goal. This is called separating the *epistemological* and the *heuristic* parts of the artificial intelligence problem and is discussed in McCarthy and Hayes (1969).

We see the following potential uses for facts about knowledge:

1. A computer program that wants to telephone someone must reason

about who knows the number. More generally, it must reason about what actions will obtain needed knowledge. Knowledge in books and computer files must be treated in a parallel way to knowledge held by persons.

2. A program must often determine that it does not know something or that someone else doesn't. This has been neglected in the usual formalizations of knowledge, and methods of proving possibility have been neglected in modal logic. Christopher Goad (to be published) has shown how to prove ignorance by proving the existence of possible worlds in which the sentence to be proved unknown is false. Presumably proving one's own ignorance is a stimulus to looking outside for the information. In competitive situations, it may be important to show that a certain course of action will leave competitors ignorant.
3. Prediction of the behaviour of others depends on determining what they believe and what they want.

It seems to me that AI applications will especially benefit from first order formalisms of the kind described above. First, many of the present problem solvers are based on first order logic. Morgan (1976) in discussing theorem proving in modal logic also translates modal logic into first order logic. Second, our formalisms leaves the syntax and semantics of statements not involving concepts entirely unchanged, so that if knowledge or wanting is only a small part of a problem, its presence doesn't affect the formalization of the other parts.

ABSTRACT LANGUAGES

The way we have treated concepts in this paper, especially when we put variables in them, suggests trying to indentify them with terms in some language. It seems to me that this can be done provided that we use a suitable notion of *abstract language*.

Ordinarily a language is identified with a set of strings of symbols taken from some alphabet. McCarthy (1963) introduces the idea of *abstract syntax*, the idea being that it doesn't matter whether sums are represented $a+b$ or $+ab$ or $ab+$ or by the integer 2^a3^b or by the LISP S-expression (PLUS A B), so long as there are predicates for deciding whether an expression is a sum and functions for forming sums from summands and functions for extracting the summands from the sum. In particular, abstract syntax facilitates defining the semantics of programming languages, and proving the properties of interpreters and compilers. From that point of view, one can refrain from specifying any concrete representation of the "expressions" of the language and consider it merely a collection of abstract synthetic and analytic functions and predicates for forming, discriminating and taking apart *abstract expressions*. However, the languages considered at that time always admitted representations as strings of symbols.

If we consider concepts as a free algebra on basic concepts, then we can regard them as strings of symbols on some alphabet if we want to, assuming that we don't object to a non-denumerable alphabet or infinitely long expressions if we want standard concepts for all the real numbers. However, if we want to regard *Equal*(X, Y) and *Equal*(Y, X) as the same concept, and hence as the same "expression" in our language, and we want to regard expressions related by renaming bound variables as denoting the same concept, then the algebra is no longer free, and regarding concepts as strings of symbols becomes awkward even if possible.

It seems better to accept the notion of *abstract language* defined by the collection of functions and predicates that form, discriminate, and extract the parts of its "expressions". In that case it would seem that concepts can be identified with expressions in an abstract language.

ACKNOWLEDGEMENTS AND BIBLIOGRAPHY

The treatment given here should be compared with that in Church (1951b) and in Morgan (1976). Church introduces what might be called a two-dimensional type structure. One dimension permits higher order functions and predicates as in the usual higher order logics. The second dimension permits concepts of concepts, etc. No examples of applications are given. It seems to me that concepts of concepts will be eventually required, but this can still be done within first order logic.

Morgan's motivation is to use first order logic theorem-proving programs to treat modal logic. He gives two approaches. The syntactic approach – which he applies only to systems without quantifiers – uses operations like our *And* to form compound propositions from elementary ones. Provability is then axiomatized in the outer logic. His semantic approach uses axiomatizations of the Kripke accessibility relation between possible worlds. It seems to me that our treatment can be used to combine both of Morgan's methods, and has two further advantages. First, concepts and individuals can be separately quantified. Second, functions from things to concepts of them permit relations between concepts of things that could not otherwise be expressed.

Although the formalism leads in almost the opposite direction, the present paper is much in the spirit of Carnap (1956). We appeal to his ontological tolerance in introducing concepts as objects, and his section on intentions for robots expresses just the attitude required for artificial intelligence applications.

We have not yet investigated the matter, but plausible axioms for necessity or knowledge expressed in terms of concepts may lead to the paradoxes discussed in Kaplan and Montague (1960) and Montague (1963). Our intention is that the paradoxes can be avoided by restricting the axioms concerning knowledge, and necessity of statements about necessity. The restrictions will be somewhat unintuitive as are the restrictions necessary to avoid the paradoxes of naive set theory.

Chee K. Yap (1977) proposes *Virtual Semantics* for intensional logics as a generalization of Carnap's individual concepts. Apart from the fact that Yap does not stay within conventional first order logic, we don't know the relation between his work and that described here.

I am indebted to Lewis Creary, Patrick Hayes, Donald Michie, Barbara Partee and Peter Suzman for discussion of a draft of this paper. Creary in particular has shown the inadequacy of the formalism for expressing all readings of the ambiguous sentence "*Pat knows that Mike knows what Joan last asserted*". There has not been time to modify the formalism to fix this inadequacy, but it seems likely that concepts of concepts are required for an adequate treatment.

REFERENCES

Carnap, R. (1956). *Meaning and Necessity*. Chicago: University of Chicago Press.

Church, A. (1951a). The need for abstract entities in semantic analysis, in "Contributions to the Analysis and Synthesis of Knowledge". *Proceedings of the American Academy of Arts and Sciences*, **80**, No. 1, 100–112. Reprinted in *The Structure of Language* (eds. Fodor, A. and Katz, J.) Prentice-Hall, 1964.

Church, A. (1951b). A formulation of the logic of sense and denotation. In *Essays in honour of Henry Sheffer* (ed. Henle, P.), pp. 3–24. New York.

Frege, G. (1892). Uber Sinn und Bedeutung. *Zeitschrift fur Philosophie und Philosphische Kritik* **100**:25–50. Translated by H. Feigl under the title "On Sense and Nominatum" in *Readings in Philosophical Analysis* (eds. Feigal, H. and Sellars, W). New York 1949. Translated by M. Black under the title "On Sense and Reference" in P. Geach and M. Black, *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: 1952.

Kaplan, D. (1969). Quantifying in, from *Words and Objections: Essays on the Work of W. V. O. Quine* (eds. Davidson, D. and Hintikka, J.), pp. 178–214. Dordrecht-Holland: D. Reidel Publishing Co. Reprinted in (Linsky 1971).

Kaplan, D. and Montague, R. (1960). A paradox regained, *Notre Dame Journal of Formal Logic*, **1**, 79–90. Reprinted in (Montague 1974).

Linsky. L. (ed.) (1971). *Reference and Modality*, Oxford Readings in Philosophy. Oxford: Oxford University Press.

McCarthy, J. (1963). Towards a mathematical science of computation, in *Proceedings of IFIP Congress* 1962. Amsterdam: North-Holland Publishing Co.

McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence* **4**, pp. 463–502 (eds. Meltzer, B. Michie, D.). Edinburgh: Edinburgh University Press.

Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability, *Acta Philosophica Fennica* **16**:153–167. Reprinted in (Montague 1974).

Montague, R. (1974). *Formal Philosophy*. New Haven: Yale University Press.

Morgan, C. G. (1976). Methods for automated theorem proving in nonclassical logics, *IEEE Transactions on Computers*, C-**25**, No. 8.

Quine, W. V. O. (1956). Quantifiers and propositional attitudes, *Journal of Philosophy*, **53**. Reprinted in (Linsky 1971).

Quine, W. V. O. (1963). *From a Logical Point of View*. New York: Harper and Row. First published Harvard University Press (1953).

Yap, Chee K. (1977). A Semantical Analysis of Intensional Logics, Research Report *RC* 6893 (#29538). Yorktown Heights, New York: IBM, Thomas J. Watson Research Center.

Foundations of a Functional Approach to Knowledge Representation

Hector J. Levesque
Fairchild Laboratory for Artificial Intelligence Research, Palo Alto, CA 94304, U.S.A.

Recommended by Terry Winograd

ABSTRACT

We present a new approach to knowledge representation where knowledge bases are characterized not in terms of the structures they use to represent knowledge, but functionally, in terms of what they can be asked or told about some domain. Starting with a representation system that can be asked questions and told facts in a full first-order logical language, we then define ask- and tell-operations over an extended language that can refer not only to the domain but to what the knowledge base knows about that domain. The major technical result is that the resulting knowledge, which now includes auto-epistemic aspects, can still be represented symbolically in first-order terms. We also consider extensions to the framework such as defaults and definitional facilities. The overall result is a formal foundation for knowledge representation which, in accordance with current principles of software design, cleanly separates functionality from implementation structure.

0. Introduction

Traditionally, the goal of knowledge representation systems has been to provide flexible ways of creating, modifying, and accessing collections of symbolic structures forming knowledge bases. In this paper, we present a new view of knowledge bases and their role in knowledge-based systems. Essentially, a knowledge base (or KB, for short) is treated as an *abstract data type* that interacts with a user or system only through a small set of operations. In our case, we consider only two interaction operations: one allows a system to ask (ASK) the KB questions about some application domain and the other allows the system to tell (TELL) the KB about that domain. The complete functionality of a KB is measured in terms of these operations; the actual mechanisms and structures it uses to maintain an evolving model of the domain are its own concern and not accessible to the rest of the knowledge-based system. In Section 1, we elaborate on this theme in terms of Newell's concept of a knowledge level.

Artificial Intelligence **23** (1984) 155–212
0004-3702/84/$3.00

From our point of view, the capabilities of a KB are strictly a function of the range of questions it can answer and assertions it can accept. In Section 2, we examine a language that can be used for questions and assertions. Specifically, we investigate in detail the semantics and proof theory of an interaction language $\mathscr{L}$ that is a dialect of first-order logic. We then show some of its limitations as an interaction language and define a new one called $\mathscr{KL}$ that greatly extends what a KB can be told or asked. This language will have the ability to refer both to the application domain and to what the KB *knows* about that domain. The operations of TELL and ASK (and consequently, the functionality of a KB) are defined in a representation-independent way in terms of this language.

In a sense, the language $\mathscr{KL}$ is the core of our functional view of a KB. First of all, the two interaction operations are defined directly in terms of the constructs it provides. But it also characterizes a KB more directly: the language contains a primitive operator $\boldsymbol{K}$ which can be read as 'the KB currently knows that . . .'. As such, the valid sentences of $\mathscr{KL}$ specify what must be known or not known by a KB of the kind we are considering.

In Section 3, we examine how to represent knowledge symbolically in order to realize the functionality required by the TELL and ASK operations. In particular, we present a proof that (under a few reasonable assumptions) the required knowledge can be represented in first-order terms even though $\mathscr{KL}$ is not a first-order language. This Representation Theorem in fact proves that a first-order implementation of a KB is *correct* in that it meets its specification (in terms of the definition of the operations).

Finally, in Section 4, we show how the framework we have presented can be extended and applied to some of the ongoing research topics in knowledge representation. Among other things, we examine how defaults (and non-monotonic reasoning) can be handled and how a 'hybrid' system containing aspects of both logical and object-oriented representation paradigms might be accomodated.

1. The Knowledge Level

In a recent paper [1], Newell introduces the idea of a *knowledge level*, an abstraction in terms of which intelligent agents can be described. For the rest of this section (and much of the paper), we will show how this idea applies equally well to knowledge bases and indeed motivates our functional approach.

1.1. Competence

A major characteristic of the knowledge level is that knowledge is considered to be a *competence* notion, "being a potential for generating action".[1] There

[1] All quotations in this section are from [1].

are actually two ways of looking at this. First, we might want to say that, to a first approximation, if an agent believes p and if p entails q, then he is likely to believe q. In this case, competence might be thought of as a plausible abstraction (or heuristic) for reasoning about the beliefs of an agent. But there is another way of looking at this: we can say that if an agent imagines the world to be one where p is true and if p entails q, then (whether or not he realizes it) he imagines the world to be one where q also happens to be true. In other words, if the world the agent believes in satisfies p, then it must also satisfy q. The point is that the notion of competence need not be one where an agent is taken to have appreciated the consequences of what he knows, but merely one where *we* examine those consequences.

So the analysis of knowledge at the knowledge level is really the study of what is *implicit* in what an agent might believe[2], precisely the domain of *logic*. As Newell points out, "just as talking of *programmerless* programming violates truth in packaging, so does talking of a *non-logical analysis* of knowledge". This is not to imply that agents do, can, or should use some form of logical calculus as a representation scheme, but only that the analysis of what is being represented is best carried out in this framework.

But what does this analysis amount to? Are there any purely logical problems to be solved? Indeed, once what Newell calls the *symbol level* has been factored out, is there *anything* concrete left to say? Newell admits the following:

> However, in terms of structure, a body of knowledge is extremely simple compared to a memory defined at lower computer system levels. There are no structural constraints to the knowledge in a body, either in capacity (i.e., the amount of knowledge) or in how the knowledge is held in the body. Indeed, there is no notion of how knowledge is held (*encoding* is a notion at the symbol level, not knowledge level). Also, there are not well-defined structural properties associated with access and augmentation. Thus, it seems preferable to avoid calling the body of knowledge a memory.

So it appears that at the knowledge level, there is nothing to say about the *structure* of these abstract bodies of knowledge called knowledge bases. This does not mean that there is nothing at all to say. Although Newell does not go into any details, he insists that "knowledge is to be characterized *functionally*, in terms of what it does, not *structurally* in terms of physical objects with particular properties and relations".

[2]The sense of 'believe' and 'know' being used here is perhaps confusing. By a 'belief', we will normally refer to something *actively* held as true. The whole point of the paper is to elucidate a sense of 'knowledge' appropriate to knowledge bases. Roughly, we intend it to refer to the logical consequences of what is believed, regardless of whether the believer is aware of the consequences or whether the beliefs are accurate.

1.2. Functionality

Newell's view of knowledge at the knowledge level is very similar to the standard notion of an abstract data type [2]: to specify what is required of a desired entity (or collection of related entities), specify the desired behaviour under a set of operations and not the structures that might be used to realize that behaviour. The canonical example is that of a *stack* that might be specified in terms of the following operations:

Create: $\rightarrow$ STACK ,
Push: STACK $\times$ INTEGER $\rightarrow$ STACK ,
Pop: STACK $\rightarrow$ STACK ,
Top: STACK $\rightarrow$ INTEGER ,
Empty: STACK $\rightarrow$ BOOLEAN .

By defining these functions (over abstract stacks), we specify *what* the behaviour should be without saying *how* it should be implemented. Of course, we have not said how stacks can be used in general to do other things (like implement recursion, for instance), but we have provided the primitives of this usage. In this sense, the specification is functional.

How might similar considerations apply to knowledge? To answer this, we have to discover the primitive operations that will be composed to form complex applications of knowledge (such as problem solving, learning, decision making, and the like). At least two orthogonal operations suggest themselves immediately.

First of all, if the behaviour of an intelligent agent is to depend on what is known, a primitive operation must be to *access* this knowledge. Very roughly speaking, we have the following:

ASK: KNOWLEDGE $\times$ QUERY $\rightarrow$ ANSWER .

In other words, we can discover answers to questions using knowledge. This is the analogue of the Top, Pop, and Empty operations which were also retrieval oriented. If we consider learning to be an intelligent activity, then we need the analogue to the Push operation: we have to be able to *augment* what is known:

TELL: KNOWLEDGE $\times$ ASSERTION $\rightarrow$ KNOWLEDGE .

Again, roughly, knowledge can be acquired by assimilating new information. There are, presumably, a large number of TELL and ASK operations corresponding to the many ways knowledge can be accessed and acquired (including non-linguistic ways). There are also other kinds of operations worth considering including an analogue of Create as well as a FORGET, an

ASSUME and others. However, for our purposes, we can restrict our attention to these two operations.

1.3. Some simplifications

To be able to say anything concrete about the operations of TELL and ASK and to simplify a lot of the machinery to come, a few assumptions are necessary. First of all, if we assume that "the knowledge attributed by the observer to the agent is knowledge about the external world", then we can assume that what a knowledge base is told or asked is also about the external world. In addition, if we now restrict ourselves to *yes–no questions*, we can assume (without loss of generality) that both the *queries* and the *assertions* are drawn from the same language. The only difference between a (yes–no) *query* and an *assertion* is that a knowledge base will be *asked* if the former is true and *told* that the latter is true.

So, for some language $\mathscr{L}$ that can be used to talk about an external world, we have the following functional specification of a KB:

$$\text{TELL: } \mathrm{KB} \times \mathscr{L} \to \mathrm{KB} ;$$
$$\text{ASK: } \mathrm{KB} \times \mathscr{L} \to \{\text{yes, no, unknown}\} .$$

What remains to be done is to settle on a suitable language $\mathscr{L}$ and come up with an abstract notion of a KB in terms of which these two operations can be defined.

It is perhaps worth examining at this stage how a *concrete* version of a KB could be realized. We might take $\mathscr{L}$ to be a dialect of the first-order predicate calculus and represent a KB as a sentence (or a finite set of sentences) of this language. In this case, TELL and ASK can be defined (in terms of first-order provability) by

$$\mathrm{Tell}(k, \alpha) = (k \wedge \alpha)$$

$$\mathrm{Ask}(k, \alpha) = \begin{cases} \text{yes} , & \text{if } \vdash (k \supset \alpha) ; \\ \text{no} , & \text{if } \vdash (k \supset \neg\alpha) ; \\ \text{unknown} , & \text{otherwise} . \end{cases}$$

There is nothing wrong with any of this, of course. This is just a standard first-order system like the kind discussed in [3] (though disguised in terms of two operations, where talk of a single query operation is much more common). However, this definition just does not address our problem any more than defining the stack operations in terms of operations on linked lists (or arrays) would. In other words, these definitions are at a different level—the symbol level—where the structure of a KB has been decided. This is not to say that there are no further design decisions to be made. In the stack case, we would

have to decide how to implement linked lists (maybe using arrays). Here, we have to decide how to realize a suitable theorem-proving behaviour over the sentences of $\mathscr{L}$ (maybe using resolution, which itself still does not say much). But the crucial choice was to *represent the knowledge symbolically* using a sentence of $\mathscr{L}$ and reduce the operations on a KB to theorem-proving operations over $\mathscr{L}$. As Newell argues,

> Logic is fundamentally a tool for analysis at the knowledge level. Logical formalisms with theorem proving can certainly be used as a representation in an intelligent agent, but it is an entirely separate issue (though one we already know much about thanks to the investigations in AI and mechanical mathematics over the last fifteen years).

Regardless of whether or not we decide to use sentences of a first-order logical language to represent what is known, a separate decision must be made about $\mathscr{L}$, the language for the TELL and ASK operations. It is the expressive power of this interface language that will determine what a KB can be said to know.[3] Since we are primarily interested in KBs whose world knowledge can be represented in first-order terms, we will start with a first-order interface language. In the next section, an appropriate first-order dialect will be defined.[4] We will then show why a more expressive *intensional* language is actually warranted. In a subsequent section, we will demonstrate that this results in abstract KB's that can still be represented in first-order terms. In sum, we will argue that a first-order language is expressively inadequate as a language for communicating with a first-order KB.

2. The Interface Language

In this section, we examine in detail some of the technical features (syntactic and semantic) a first-order dialect should have to be useful as the interface language $\mathscr{L}$. We will then show that, in fact, a language that is more powerful than this first-order language is required. This will lead to a new interface language $\mathscr{KL}$ whose use will be explained and whose semantics will be defined (by extending the semantics of $\mathscr{L}$). Finally, the TELL and ASK operations over this language will be defined.

[3]The situation is complicated here by the fact that the TELL and ASK operations may not allow precisely the same subset of the interface language to be used. As an extreme case, TELL may be restricted to atomic sentences (as in relational style databases) while ASK may allow a full first order query language. For example, you might be able to find out if $(p \vee q)$ is true but only be able to inform the KB that p is true or that q is true.

[4]The fact that we want to characterize what a KB knows at the knowledge level other than by a collection of symbolic structures will force us to use the *semantics* of $\mathscr{L}$ directly in our definitions. While this may present a problem for those readers more familiar with symbolic or *proof-theoretic* uses of formal languages, we have, at least, designed the semantic theory of $\mathscr{L}$ with the specific goal of TELL and ASK in mind.

2.1. A first-order language

As pointed out in [4], it is one thing to say that *some* first-order language is going to be used for some purpose but quite another matter to say *which*. A traditional use of logical languages in AI has been to *represent* knowledge symbolically and has lead to considerations such as prenex forms, Skolemization and various theorem-proving strategies. Our demands are quite different in that we seek a logical language $\mathscr{L}$ as a medium for querying and updating a KB.

2.1.1. *Singular terms*

The first and perhaps most radical characteristic of the language $\mathscr{L}$ is its treatment of *singular terms*. While there is a sense in which constant and function symbols are theoretically dispensable in first-order settings (see, for example, [5, pp. 84–85]), in our case, some form of singular term is necessary. Specifically, if we ever want to be able to ask a KB a *wh-question* and obtain an answer other than *unknown*, we must be able to provide the information necessary to answer the question. We have to not only be able to tell the KB that there exists *someone* with a certain property; we must be able to tell it *who* that individual is and have the KB realize that it is being told precisely that.

For example, if we consider telling a KB that[5]

CoastalCity(LargestCity(USA)),

we must be clear as to whether we are trying to tell it *what city* is a coastal city or merely (and more plausibly) *that some city* (that also happens to be the largest American city) is. But under what condition do we want to say that the KB knows what the largest American city is? It certainly does not amount to simply knowing something like

(LargestCity(USA) = Birthplace(EdwardMcDowell))

unless the KB also knows where McDowell was born. But then we might argue that even if the KB knows that

(LargestCity(USA) = NewYorkCity),

it still does not know the largest city unless it knows what city New York City is. To allow a KB to answer wh-questions, we have to *design* $\mathscr{L}$ in such a way as to avoid this potentially infinite regress.

[5]Here and elsewhere, the reverse PROLOG convention will be used; variables will appear in lower case, while function, constant and predicate symbols will begin in upper case.

One way to do this is to consider the entire space of singular terms as being partitioned into equivalence classes where, intuitively, two terms are in the same class if they corefer. Then we might say that a KB knows what the largest city is if it knows what equivalence class the term belongs to. To inform the KB of this, we need only notice that since there are countably many singular terms in $\mathscr{L}$, there can be at most countably many equivalence classes. So we can simply introduce into $\mathscr{L}$ a countable number of new terms (which we call *parameters*)

$$1, 2, 3, \ldots$$

understood as distinct but semantically vacuous representatives of each equivalence class.[6] By convention, then, to know what equivalence class a term t belongs to (and therefore be able to answer the wh-question) is to know something of the form

$$(t = k) \quad \text{for some parameter } k\,.$$

So, for example, if a KB knows that

$$(\text{NewYorkCity} = \text{TheBigApple}) \wedge (\text{TheBigApple} = 2)\,,$$

then we will say it knows what city New York and The Big Apple are. Of course, there is at least one sense in which this is contentious or just plain false—knowing what city New York is *should* involve New York somehow. But recall that the sense of 'know' we have in mind is a conventional one where first, it makes sense for a KB to have knowledge at all and second, this depends on what questions it can answer, not on what connections it has to the world. The issue here is not how well acquainted a KB is with New York City, but rather when a certain wh-question can be answered.

2.1.2. *Incomplete knowledge*

Given the above discussion of wh-questions, the only other constraint that plays a role in determining the form of $\mathscr{L}$ is the *incompleteness* of knowledge bases. As argued in [6, 7], to allow a knowledge-based system to capitalize on whatever knowledge about its application domain is available, it must be possible for the KB to find out about the world in an incremental way. The only alternative is to postpone the assimilation of information until it has become sufficiently specific which, for many applications, may simply *never*

[6]Parameters are thus *standard names* of members of the equivalence classes. Alternately, we could introduce some special way of stating for each number i and term t that t belongs to the ith equivalence class.

happen. In terms of TELL and ASK, what this amounts to is that $\mathscr{L}$ has to be sufficiently expressive to allow us to make weak assertions about the world.

Consider, for example, the sentence[7]

$$\text{LosAngeles} \neq \text{Capital(California)}.$$

The question is why (apart from convenience) we would ever need a sentence like this in $\mathscr{L}$. Why not instead just tell the KB what the capital of California is:

$$\text{Capital(California)} = \text{Sacramento}$$

(assuming it already knows that Sacramento and Los Angeles are distinct cities)? From this sentence, it is a simple matter to infer the negated sentence. Similarly, why would we ever bother telling the KB that

$$(\text{Capital(California)} = \text{SanFrancisco}) \vee (\text{Capital(California)} = \text{Sacramento})$$

since the disjunction can be easily inferred once we tell it what the capital is? The point should be clear: in terms of communicating with a KB, it's not so much that negation and disjunction allow information to be conveyed that otherwise could not, but only that they do so *without also requiring more information* than what may be available (such as knowing what the capital is). In other words, they allow a KB to be told *exactly* what is known about the world, however vague. If more specific information ever becomes available, then it will be assimilated as well, but until that point, the KB can at least use whatever incomplete knowledge it has acquired.

Much the same argument can be made for the inclusion in $\mathscr{L}$ of existential quantification and non-parameter terms (which need not corefer). The conclusion to be drawn is that the standard features of a first-order language such as non-parameter terms, negation, disjunction and existential quantification, can all be motivated from the point of view of incomplete knowledge bases; they all allow knowledge to be acquired at the appropriate level of vagueness or generality.[8] If the source of information was suitably rich, the expressive power would not be necessary (though obviously convenient, in the case of universal quantification).

[7]The term 'sentence' is being used technically here to mean a closed formula of $\mathscr{L}$.

[8]Interestingly enough, second- (or ω) order quantification (over properties, relations and the like) is much more difficult to motivate in these terms. While it makes sense to want to be able to say that every major city of New York has a certain property (without having to know what cities are the major ones) it seems strange to want to say anything about *every property* of New York (including the negation of every property it does not have).

2.1.3. *Semantics*

The language $\mathscr{L}$ we are considering is built up in the obvious recursive way from a set of function and predicate symbols of every arity (including constants, the 0-ary function symbols), parameters (which behave syntactically exactly like constants), (individual) variables, and the usual logical punctuation. The function symbols (including the constants) and the predicate symbols (excluding the equality symbol) are considered to be the *non-logical* symbols; all others are *logical* symbols. Expressions in the language are syntactically divided into *terms* and *sentences*: the terms are variables, parameters and function applications; the sentences are predicate applications, equality expressions, negations, disjunctions and quantifications. Sometimes disjunction will be used as *the* binary propositional connective and, other times, material implication. Similarly, existential or universal quantification will be *the* quantifier. The intent is that any other connective (such as material equivalence) be made available in the usual way. For any variable x, closed term t and expression α (term or sentence), the expression α_t^x is the result of replacing all free x in α by t. Finally, if $x_1, x_2, \ldots, x_n$ are the free variables in α, then $\alpha[t_1, t_2, \ldots, t_n]$ is the closed expression resulting from simultaneously replacing each free x_i by t_i.

We will describe the semantics of this language using what Quine has called *truth value semantics*.[9] While this form of semantic specification does not shed any light on the referential nature of the language, it forms a more convenient semantic basis for the language $\mathscr{KL}$, to be introduced later. Essentially, the idea is to specify (with as little fuss as possible) which of all the possible mappings from sentences of $\mathscr{L}$ to true or false are admissible interpretations.

Consider, for example, an assignment of truth values to sentences of the form $t_1 = t_2$. The kind of thing we obviously want to rule out is an assignment where, say, the three sentences

$$f(1) = 2, \qquad c = 1 \quad \text{and} \quad f(c) = 3$$

are all interpreted as true.

The easiest way to do this is to make the assignment to equality sentences depend on a partitioning of the primitive terms of the language. The *primitive terms* are those that contain exactly one function symbol. So, for example,

$$c, \qquad f(1) \quad \text{and} \quad g(2, 1)$$

are all primitive terms; if these all belong to the same equivalence class as the parameter *1*, then so must the non-primitive terms

$$f(c), \qquad f(g(2, c)) \quad \text{and} \quad g(2, f(f(1))).$$

[9]See [8], for an introduction to (and a defense of) this kind of semantic theory.

To enforce this constraint, we first assign a parameter (or equivalence class) to each primitive term. Given v, an assignment of this kind, we can define what it means for any two terms of $\mathscr{L}$ to corefer as follows:

The *coreference relation* (given v) is the least set of pairs such that:

(1) if t is a primitive term, then t and $v[\![t]\!]$ corefer;

(2) if t_1 and t_2 corefer, then so do $t^x_{t_1}$ and $t^x_{t_2}$.

It is a simple matter to show that for each v, this assigns a unique coreferring parameter (and thus, a unique equivalence class) to each term of the language. In particular, this definition of coreference guarantees that *1* and *2* do not corefer on any v.

An assignment of truth values to sentences of $\mathscr{L}$ is much like an assignment to terms of $\mathscr{L}$ except that there are only two possible equivalence classes: the true sentences and the false ones. As above, the easiest way to rule out inadmissible interpretations is to make them depend on assignments to the primitive sentences (and for equality sentences, to the primitive terms). The *primitive sentences* are the atomic ones that contain no function symbols (so all primitive expressions, terms or sentences, have exactly one non-logical symbol). Given a set s of primitive sentences (taken to be the true ones) and an assignment v as above (taken to specify coreferentiality), the truth value of every sentence of $\mathscr{L}$ can be defined as follows:

The *true sentences* on s and v form the least set such that

(1) every element of s is true on s and v;

(2) if t_1 and t_2 corefer given v, then $(t_1 = t_2)$ is true on s and v and the atomic sentences $\rho^x_{t_1}$ and $\rho^x_{t_2}$ have the same truth value on s and v;

(3) if α is not true on s and v, then $\neg\alpha$ is;

(4) if either α or β is true on s and v, then so is $(\alpha \vee \beta)$;

(5) if for some parameter i the sentence α^x_i is true on s and v, then so is $\exists x\alpha$.

The definition can be extended in the usual way to include conjunctions, implications, equivalences and universal quantifications based on negations, disjunctions and existential quantifications. Given this definition of truth, we define a sentence as *valid* if it is true on every s and v. A set of sentences is *satisfiable* if there is an s and v on which every element of the set is true.

The first thing to notice about the above definition is that when we restrict our attention to that subset of $\mathscr{L}$ without parameters or equality, it produces the same valid and satisfiable sentences as the standard one due to Tarski. As for equality, to the extent that it is considered to be part of a first-order language at all, it is usually treated as a regular predicate with a set of axioms in a theory forcing the predicate to be an equivalence relation. Semantically, as far as the language is concerned, this leaves the predicate symbol uninterpreted. In our case, however, the language is constrained to have a very definite interpretation of equality. Specifically, once an assignment of equivalence classes to terms has been made, the interpretation of equality sentences is fixed.

One major difference between our definition of validity and a more standard one is that in our case only, the following sentences are all valid:

$$\begin{aligned}
&\neg\exists x_1 \forall y(y = x_1),\\
&\neg\exists x_1 \exists x_2 \forall y(y = x_1) \vee (y = x_2),\\
&\neg\exists x_1 \exists x_2 \exists x_3 \forall y(y = x_1) \vee (y = x_2) \vee (y = x_3),\\
&\quad\vdots
\end{aligned}$$

In Tarskian terms, what this amounts to is a restriction on interpretations forcing the domain of quantification to be infinite. In practice, this is not a very serious restriction since in no way does it constrain the interpretation of the predicate (and function) symbols, all of which may be taken to have finite or infinite extensions. It is only sentences that talk about the totality of what exists that will have different interpretations. These sentences are indeed special in our case since they contain only logical symbols and so are either valid or unsatisfiable. Another way of looking at this is to say that there is no real way in $\mathscr{L}$ to talk about what exists as a whole except by talking about the extensions of the predicate or function symbols. This suggests that a *typed* form of quantification might be more compatible with this semantics. This would syntactically rule out the above list of valid sentences which talk about the size of the entire domain.

One final point to observe about the semantics of $\mathscr{L}$ is that the Compactness Theorem fails for this dialect. In other words, just because every finite subset of a set of sentences is satisfiable does not mean that the set itself is. An example of such a set is $\{\exists x\phi(x), \neg\phi(1), \neg\phi(2), \neg\phi(3), \ldots\}$ for some monadic predicate ϕ. Again this has little practical consequence since our major concern will be with sentences and finite sets of sentences of $\mathscr{L}$.

So, to summarize, the truth theory of $\mathscr{L}$ is based on a partition of the primitive sentences into two groups and a partition of the primitive terms into an infinite number of groups. Together, the partitions determine the truth value of every sentence in the language. Thinking of the language $\mathscr{L}$ as talking about some (possible) world (or world state, situation, state of affairs, slice of reality), an s and a v together tell us exactly what that world is like in as much detail as the language $\mathscr{L}$ allows. We will consequently call a pair $\langle s, v\rangle$ a *world structure* since it specifies a world relative to the language $\mathscr{L}$. Of course, many worlds can correspond to such a structure, namely those whose differences are not captured by the language $\mathscr{L}$.

2.1.4. *Proof theory*

The semantics for $\mathscr{L}$ presented above does not require dealing with formulas having free variables. The proof theory will also have this property in that at

any stage of a proof, only *sentences* of $\mathscr{L}$ will be considered, with parameters playing the role of free variables.

The major deviation from standard axiomatizations of first-order logic involves the relationship between quantification and parameters. Specifically, to guarantee soundness and completeness (with the above semantics), we must insure that

$$\forall x\alpha \text{ is a theorem} \quad \text{iff} \quad \text{for every parameter } k,\ \alpha^x_k \text{ is a theorem}.$$

The 'only if' part of this is easy to guarantee with an axiom schema (usually called the Axiom of Specialization)

$$\forall x\alpha \supset \alpha^x_t \quad \text{for any term } t.$$

The modification to the rule of Universal Generalization necessary to handle the converse is a bit more troublesome. The obvious 'rule' is

$$\text{From } \alpha^x_1, \alpha^x_2, \ldots, \alpha^x_i, \ldots \text{ infer } \forall x\alpha$$

which unfortunately leads to a notion of proof that requires an *infinite* number of subproofs before a universal can be concluded. On the other hand, one might be tempted to reason as follows:

> Suppose I am able to prove a theorem α which contains some parameter k. There is nothing special about k, so I should be able to also prove the theorem for any other parameter. Consequently, I am justified in concluding that $\forall x\alpha'$ where α' is α with k replaced by x.

Except for the equality predicate, this reasoning is sound. The trouble with equality is that the parameter k is indeed special in that it is the only parameter that is equal to k and to no others. In particular, we do not want to be able to reason

$$\ldots 1 \neq 2, \text{ and therefore, } \forall x(x \neq 2)\ldots$$

since *1* is special in not being equal to *2*. So the reasoning in general has to be as follows:

> Suppose I am able to prove α which contains parameter k as well as parameters $i_1, \ldots, i_n$. If I am also able to prove α but with k replaced by $i_1, \ldots, i_n$, then there is no way the theorem could have been a consequence of the fact k is special relative to the other i_j. Thus, I can conclude that $\forall x\alpha'$.

As it turns out, *this* reasoning is sound[10] and leads to the following version of Universal Generalization:

> From $\alpha^x_{i_1}, \ldots, \alpha^x_{i_n}$, infer $\forall x\alpha$ provided the i_j range over all the parameters in α and at least one not in α.

The only other aspect of the proof theory of $\mathscr{L}$ to worry about is the treatment of equality itself. Fortunately, the parameters permit a very simple and elegant formalization. All that is required is a single axiom schema (called the Axiom of Equality) to state that only identical parameters are equal:

$$(i = i) \wedge (i \neq j) \quad \text{for any two distinct parameters } i \text{ and } j.$$

The fact that this axiom schema has a proviso attached is of no real consequence. Standard formulations of first-order logic also attach a proviso to the Axiom of Specialization to avoid the collision of quantifiers. In [7], it is shown that all the usual properties of equality (like the Leibniz property) can be derived from this axiom. For example, we can prove that equality is a symmetric relation by Universal Generalization from

$$\forall y(1 = y) \supset (y = 1),$$

which derives from

$$(1 = 1) \supset (1 = 1) \quad \text{and} \quad (1 = 2) \supset (2 = 1),$$

both direct consequences of the Axiom of Equality. So, to summarize, we have the following proof theory for $\mathscr{L}$:

Axioms.

(A1) $\alpha \supset (\beta \supset \alpha)$.
(A2) $(\alpha \supset (\beta \supset \gamma)) \supset ((\alpha \supset \beta) \supset (\alpha \supset \gamma))$.
(A3) $(\neg\beta \supset \neg\alpha)) \supset ((\neg\beta \supset \alpha) \supset \beta)$.
(AD) $\forall x(\alpha \supset \beta) \supset (\forall x\alpha \supset \forall x\beta)$.
(AS) $\forall x\alpha \supset \alpha^x_t$.
(AE) $(i = i) \wedge (i \neq j)$ *for all distinct i, j.*

[10]See [7] for details. Note also that the reasoning remains sound when 'provable' is replaced above by 'valid'. However, it is unsound when 'provable' is replaced by 'true'. To conclude that $\forall x\alpha$ is true, we would need *all* instances of α^x_k as evidence.

Rules of inference.

(MP) *From* α *and* $\alpha \supset \beta$, *infer* β.

(UG) *From* $\alpha^x_{i_1}, \ldots, \alpha^x_{i_n}$ *where the* i_j *are those in* α *and one not in* α, *infer* $\forall x \alpha$.

This gives us a precise notion of theoremhood in $\mathscr{L}$ that is extensionally equivalent to the above definition of validity. That is, it can be shown that

$$\vdash \alpha \quad \text{iff} \quad \alpha \text{ is valid}.$$

So it should be no surprise that the theorems (at least those without parameters and equality) correspond to those of the standard axiomatization.

2.2. An extended interface language

The above definition of validity and theoremhood can be used to define the two interface functions TELL and ASK that would use $\mathscr{L}$ to link a KB to an external agent (or program) that is using the KB as a repository for information about some application domain. However, we will argue that the language $\mathscr{L}$ is an appropriate query language for ASK only if the language for TELL is limited in its use of negation, disjunction and existential quantification. The expressive power of $\mathscr{L}$ as an assertion language simply makes it too weak as a query language.

2.2.1. *First-order limitations*

Consider, for example, a KB that is keeping track of state capitals:

Capital(Pennsylvania) = Harrisburg
Capital(Indiana) = Indianapolis
.
.
.

Conceptually, at least, there is certainly no problem with a wh-question

Capital(state) = city?

which asks the KB to list the states and their capitals (by using free variables *state* and *city*). This is in fact what would happen with a PROLOG data base and query similar to the above. The trouble arises when we consider a KB that has much less knowledge about this domain. For example, a KB may only know

Capital(Texas) = Houston ∨ Capital(Texas) = Austin
Capital(California) ≠ LosAngeles

without having identified the capital of either state.[11] In this case, the request to list the states and their capitals is much more problematic. In general (and at best), the system will be able to provide some *description* of the capitals based on what it knows such as "The capital of California is not a coastal city". But equally *true* would be to say "The capital of California is a city that is not located in Wyoming nor in New York". To avoid *that* kind of answer, some measure of relevance would have to be established based, presumably, on a theory of what an acceptable answer in this domain should be and what the user already knows. One can imagine, for example, a system engaging in a clarification dialogue with the user about an acceptable form of answer. In other words, this kind of ability for wh-questions goes beyond a simple interface to a KB and starts looking more and more like a separate knowledge-based system with a KB of its own. There is, in fact, a minor problem even in the first case if terms like 'Harrisburg' in the above are really *constants*, not parameters. For example, a KB could contain $\phi(c)$ for a large number of constants c while still knowing that there was a unique ϕ. Once again, it is not clear what the answer to a wh-question should be.

To the extent that a KB is intended to provide a service to a knowledge-based system, it should, at the very least, be able to inform a user about when a wh-question is problematic. From there, it would be a separate step to decide what to do about an answer. For example, consider the question that asks for the capital of California:

$$\text{Capital(California)} = \text{city}\,.$$

This question has a direct answer only when the system knows what the capital is. To find out if the KB has that knowledge, the question

$$\exists\text{city}[\text{Capital(California)} = \text{city}]$$

is not adequate since this asks if California has a capital and not if the system knows what it is. In other words, the kind of question we have to ask is not whether or not there is something that *is* the capital but whether or not there is something that *the* KB *believes to be* the capital. The difference is crucial: the first question asks the KB about the *world*, whether or not California has a capital; the second question asks the KB about the KB *itself*, whether or not it has on record the capital of California.

If we accept the position that the language $\mathscr{L}$ is intended to talk about the world only, then we have to go beyond $\mathscr{L}$ to be able to ask questions that also

[11]This is, of course, precisely the kind of knowledge that *cannot* be represented in the subset of first-order logic embodied in PROLOG. Whatever its merit as a programming formalism, PROLOG is severely limited as a representation language.

talk about the KB. The interface language $\mathcal{KL}$ contains all of $\mathcal{L}$ and, in addition, for any formula α contains a formula $\boldsymbol{K}\alpha$. A sentence $\boldsymbol{K}\alpha$ here should be read as "The KB currently knows that α". Among the differences between the two sentences, note that if the question α (i.e., is α true?) is answered *unknown*, then the question $\boldsymbol{K}\alpha$ (i.e., is α known?) should be answered *no*. Before describing the exact semantics of the language $\mathcal{KL}$, it perhaps worth examining how it can be used as an interface language.

2.2.2. *The use of* $\mathcal{KL}$

The most important property of $\mathcal{KL}$ is that whereas we can use a question like

$$\exists \text{city}[\text{Capital}(\text{California}) = \text{city}]$$

to find out if California has a capital, we can now also use

$$\exists \text{city}\boldsymbol{K}[\text{Capital}(\text{California}) = \text{city}]$$

to find out if the KB knows any of them. In other words, the KB will answer *yes* to the first one if it thinks California has a capital and to the second one if it thinks it knows one of them. The first sentence is true of a *world* where California has a capital; the second is true of a *knowledge base* that knows a capital of California. Similarly, the sentence

$$\exists \text{state} \neg \exists \text{city}\boldsymbol{K}[\text{Capital}(\text{state}) = \text{city}]$$

is true of a KB when there is a state such that no city is known by the KB to be its capital.

There are also sentences in $\mathcal{KL}$ that talk simultaneously about the world and the KB. The sentence

$$\exists \text{city}[\text{MajorCity}(\text{city}, \text{California}) \wedge \neg \boldsymbol{K}\text{MajorCity}(\text{city}, \text{California})]$$

will be true if California really has a major city that the KB does not know to be a major city. For this sentence to be true, the world and the KB have to be in a certain state. If asked as a question, we are asking the KB for an opinion on the relationship between what is true and what it knows. The answer can be *yes*, *no* or even *unknown*. For example, if the KB knows that

$$\forall \text{city}[\text{MajorCity}(\text{city}, \text{California}) \equiv (\text{city} = \text{LosAngeles})],$$

then, assuming the KB knows what city LosAngeles is, the answer would be *no* since the KB knows there can be no major city of California that it does not

know about. Of course, the KB can be mistaken about California and major cities, but that is a separate issue. Given the above belief, the KB is certainly justified in believing that it knows all the major cities of California. On the other hand, if the KB only knows that

$$\text{MajorCity(Fresno, California)} \vee \text{MajorCity(Livermore, California)},$$

then the answer to the question is *yes* since the KB knows that California has a major city that is not among the known ones. Although it does not know *what* city has that property, it does know that there is one, and moreover, it is either Fresno or Livermore. The reasoning here, from the point of view of the KB, is somewhat subtle. The KB knows that neither Fresno nor Livermore are known to be major cities. Moreover, it knows that one of them is a major city. Thus, it knows that there is a major city that is not known to be a major city. Finally, in the simplest case where the KB has no information at all about the major cities, the answer is *unknown* since the KB has no way of knowing whether or not California has major cities it might not know about.

The above sentence can be generalized in an interesting way. We can first abstract from California in the above and get

$$\exists \text{state} \neg[\exists \text{city}(\text{MajorCity}(\text{city}, \text{state}) \wedge \neg \boldsymbol{K}\text{MajorCity}(\text{city}, \text{state}))] .$$

This sentence will be true if, for some state, the KB knows all of its major cities. On the other hand, the truth of the sentence

$$\exists \text{state} \boldsymbol{K} \neg[\exists \text{city}(\text{Majorcity}(\text{city}, \text{state}) \wedge \neg \boldsymbol{K}\text{MajorCity}(\text{city}, \text{state}))]$$

once again depends only on the KB and not on the world. Roughly speaking, this sentence will be true if the KB knows a state for which it thinks it knows all the major cities.[12] This sentence talks about the *meta-knowledge* of a KB, that is, the knowledge it has about what it knows about the world. This can be seen in the syntax of the sentence itself where a $\boldsymbol{K}$-operator occurs within the scope of another $\boldsymbol{K}$-operator.

So far, we have concentrated on how a language like $\mathcal{KL}$ can be used as a query language to ask the KB not only about what it thinks is *true*, but what it thinks is *known*. But $\mathcal{KL}$ can also be used to *tell* a KB about the world by referring to what is already known. To see this, we must first examine how $\mathcal{KL}$ *cannot* be used as an assertion language.

Consider, for example, a sentence of $\mathcal{KL}$ of the form $(\boldsymbol{K}\alpha \vee \boldsymbol{K}\neg\alpha)$. Intuitively, this sentence is true of a KB when that KB either knows α or knows

[12] More precisely, the sentence is true if there is something (say, a state) that has the property that the KB thinks that there is no city that is a major city of that state without the KB knowing it.

$\neg\alpha$. As a question to a KB, the answer to this sentence should be *yes* if the KB knows whether α is true and *no* otherwise. Suppose we now consider telling a KB that this sentence is true. If the KB already knows the truth value of α, the sentence does not tell the KB anything about itself that it does not already know and so the assertion is redundant. If, on the other hand, the KB does not know the truth value of α, then telling it the sentence is certainly not going to change anything since it will not help the KB decide what the truth value is. In fact, the KB already believes the sentence to be false and the only way it can (and should) change its mind is when it discovers the truth value of α. In this case, then, the assertion of the sentence contradicts what is already known. A KB cannot be informed of the truth value of α simply by being told that it knows the truth value. Such an assertion will either be redundant or contradictory.

The problem with asserting $(\boldsymbol{K}\alpha \vee \boldsymbol{K}\neg\alpha)$ is that this sentence does not say anything at all about the world. It is true or false based only on the current state of the KB. Its truth, therefore, cannot possibly affect the picture of the world that the KB might have. But there are sentences of $\mathcal{KL}$ (over and above those of $\mathcal{L}$) that have this property. For example, the sentence

$$\forall \text{city}[\text{MajorCity}(\text{city}, \text{Wyoming}) \supset \boldsymbol{K}\text{MajorCity}(\text{city}, \text{Wyoming})]$$

is true when the KB knows every major city of Wyoming. As a question, the KB will answer *yes* if it thinks it knows every major city, *no* if it thinks it does not and *unknown* otherwise. An instance of this last case is when all the KB knows is that

$$\text{MajorCity}(\text{Cheyenne}, \text{Wyoming}).$$

In this case, it *does* make sense to inform the KB that it knows all the major cities since, in a roundabout way, this tells it something about the world, namely that Cheyenne is the *only* major city of Wyoming. As discussed in [6, 9], this is an instance of the *closed world assumption*, relativized to the major cities of Wyoming. The sentence in fact tells the KB that any city not *known to be* a major city of Wyoming *is not* a major city of Wyoming. Similarly, telling a KB that

$$\forall \text{city}[\exists \text{state}\ \text{MajorCity}(\text{city}, \text{state}) \equiv \exists \text{number}\boldsymbol{K}(\text{Population}(\text{city}) = \text{number})]$$

tells it that the major cities are exactly those cities whose populations are (currently) known. Finally, telling a KB that

$$\exists \text{city}[\text{MajorCity}(\text{city}, \text{NewYork}) \wedge \neg\boldsymbol{K}\text{MajorCity}(\text{city}, \text{NewYork})]$$

is a way of asserting something about New York, that is, that it has a major city that the KB does not know about.

Given a more general interface language like $\mathscr{KL}$, the definition of TELL and ASK must be reconsidered. Specifically, we can no longer use the intuition that a question α must be answered *yes* when it is a (first-order) consequence of what is in the KB, since α now may contain $\boldsymbol{K}$ operators. However, we do know that α should be answered *yes* when it is known to be true, that is, precisely when $\boldsymbol{K}\alpha$ is true. So the answer to questions in this more general framework is determined by the truth conditions of the language $\mathscr{KL}$, which we will now examine.

2.2.3. *Competence and closure*

The semantics of $\mathscr{KL}$ is based on two assumptions about the sense of knowledge we are considering. The first assumption, which was motivated above, is the Assumption of Competence. The second assumption, which deals with meta-knowledge only, is called the Assumption of Closure. In this section, we examine these two assumptions in preparation for a truth theory of $\mathscr{KL}$ based on the one for $\mathscr{L}$.

The first assumption might be phrased as follows:

Assumption of Competence. *Every logical consequence of what is known is also known.*

Again this is not to say that a KB actively believing α and $(\alpha \supset \beta)$ must somehow be *logically forced* into actively believing β. Rather (in the case of sentences of $\mathscr{L}$ anyway), it says that if a KB has a picture of a world where both α and $(\alpha \supset \beta)$ are true, then the world it imagines must be one satisfying β.[13] This assumption places a lower limit on what knowledge a KB can have since it forces every KB to know at least the valid sentences and the consequences of what it knows.

An upper limit on what a KB can know is also determined by this assumption. If a KB knows both α and $\neg\alpha$, then *every* sentence of $\mathscr{KL}$ must be known. In other words, any KB that has inconsistent knowledge must know exactly the same sentences. It is important to realize that there is absolutely nothing contradictory about a KB that is inconsistent. Nor is it a defect of logic that an inconsistent KB knows every sentence. It is just a property of our definition of negation in $\mathscr{L}$ that a sentence and its negation cannot be simultaneously satisfied. What is implicit in a picture of a world where both a

[13]This is simply a consequence of the truth conditions of the language. There is no notion of the KB *doing* anything here (like a proof, for instance), but only us examining the consequences of what it knows.

sentence (of $\mathscr{L}$) and its negation are true? It is simply the case that in every world so described *of which there can be none*, any sentence of $\mathscr{L}$ is true. As it turns out, however, the Assumption of Closure described below will constrain a KB to be *consistent*, so the point is really moot.

This is the only assumption about world knowledge that will be made. Specifically, there is no assumption that the world knowledge of a KB is *complete*. In other words, there may be sentences of $\mathscr{L}$ that are neither known to be true nor known to be false. In addition, there is no assumption that the world knowledge is *accurate*. Just because the KB thinks that something is true does not mean that it *really* is. While ideally it would be nice to have a KB that was a complete and accurate model of reality, the well-formedness of a KB does not depend on these two properties.

One way of looking at the competence of a KB semantically is to think of a KB as an incomplete picture of some world, a world where what the KB knows is true. In general, there will be many distinct worlds that satisfy everything that the KB knows.[14] We can call these worlds *compatible* with the KB. For example, if a KB knows α and $(\sigma \vee \beta)$, then a world satisfying α, $\neg\beta$, and σ is compatible with the KB but one satisfying $\neg\alpha$ is not. Given this characterization of compatible worlds, the Assumption of Competence tells us that a sentence of $\mathscr{L}$ is known exactly when it is true in all compatible worlds. So we can identify in an abstract way the *world knowledge* held by a KB with its set of compatible worlds. In other words, if we want to treat the knowledge held by a KB as a distinct abstract entity, a set of worlds seems to be a reasonable abstraction to use. We can then use a *set* of world structures to specify a KB as anything having this world knowledge.

This set theoretic view of knowledge bases gives us a natural picture of knowledge acquisition. One can imagine starting with a KB that has no world knowledge and so is described by the set of *all* world structures. If it now finds out that α is true, any world where α is false is no longer compatible with the KB and so the KB is described by a smaller set of world structures. As more and more knowledge is acquired, more and more worlds are eliminated. In the limit, one can imagine a KB knowing the truth value of every sentence of $\mathscr{L}$ and thus having narrowed its picture of the world down to the point where only the worlds corresponding to a *single* world structure satisfy what it knows. This is as far as we can go using $\mathscr{L}$. Any additional non-redundant information leads to an inconsistent KB where the set of compatible worlds is empty.

If we are assuming that a KB can address any question phrased in $\mathscr{KL}$, then we are considering a KB with meta-knowledge as well as world knowledge. In fact, implicit in the above discussion was that a KB could answer questions about its own knowledge accurately and even completely. If we call a sentence

[14]This might be true even if the KB is *complete*, that is, knows the truth value of every sentence of $\mathscr{L}$, since there may be aspects of worlds that $\mathscr{L}$ does not address.

of $\mathcal{KL}$ *pure* if it is just about the knowledge of the KB (that is, is true or false independently of the world),[15] then we have the final assumption:

Assumption of Closure. *A pure sentence is true exactly when it is known.*

This assumption states that no matter how incomplete or inaccurate the *world* knowledge of a KB may be, it always has complete and accurate knowledge about its own knowledge. Of course, the assumption has to be understood in the context of the competence assumption. The KB need not actually derive all true sentences about itself. The knowledge is only *implicitly* there. On the other hand (and as discussed above), the assumption is that there will never be any reason to inform the KB about itself; we can concentrate on providing it world knowledge and leave it the responsability of keeping its meta-knowledge up to date. But more importantly, there is never any reason to doubt what a KB knows about itself; it is the final authority on what it does and does not know.

The last property is enough to guarantee the *consistency* of a KB. The reason is that an inconsistent KB would have to know every sentence including one saying that it did not know another. But, by the closure assumption, it would have to be true that it did not know the latter sentence, contradicting the assumption that it knows every sentence.

The two assumptions together tell us how to interpret any sentence $\boldsymbol{K}\alpha$ given the knowledge of a KB characterized as a set of worlds. If α is pure, then $\boldsymbol{K}\alpha$ will be true whenever α is. If α is a sentence of $\mathcal{L}$, $\boldsymbol{K}\alpha$ will be true iff every compatible world satisfies α. Finally the case where α is neither pure nor first-order will be handled by appealing to the first two cases.

2.2.4. *The semantics of* $\mathcal{KL}$

Given this informal interpretation of the $\boldsymbol{K}$ operator, we can present a truth value semantics for $\mathcal{KL}$ similar to the one presented earlier for $\mathcal{L}$. Since $\mathcal{KL}$ includes all of $\mathcal{L}$ as a special case, an interpretation of the sentences of $\mathcal{KL}$ depends on both a world and a KB. Thus we require a specification of both a world and a KB to determine the truth value of every sentence. As before, we can take a world structure to be a pair $\langle s, v \rangle$ where s is a set of primitive sentences and v is an assignment of parameters to primitive terms. A KB *structure* can be taken to be any non-empty set of world structures, intuitively, those that are compatible with the KB.[16] If k is a KB structure, then given a world structure consisting of s and v, we can define the truth value of any sentence of $\mathcal{KL}$ as follows:

[15]Syntactically, a sentence of $\mathcal{KL}$ is *pure* iff every predicate symbol (excluding equality) and every function symbol (including constants) appears within the scope of a $\boldsymbol{K}$.

[16]As we will discuss later, not every set of world structures is appropriate here.

The *true sentences* on s, v and k form the least set such that:

(1) every element of s is true on s, v and k;

(2) if t_1 and t_2 corefer given v, then $(t_1 = t_2)$ is true on s, v and k and the atomic sentences $\rho^x_{t_1}$ and $\rho^x_{t_2}$ have the same truth value on s, v and k;

(3) if α is not true on s, v and k, then $\neg\alpha$ is;

(4) if either α or β is true on s, v and k, then so is $(\alpha \vee \beta)$;

(5) if for some parameter i the sentence α^x_i is true on s, v and k, then so is $\exists x\alpha$.

(6) if α is true on s', v' and k for every $\langle s', v'\rangle$ in k, then $\boldsymbol{K}\alpha$ is true on s, v and k.

This definition duplicates the one for $\mathcal{L}$ except for the last clause which stipulates that $\boldsymbol{K}\alpha$ is true on k iff α is true on every element of k (and k itself, if α is not first-order). If α is first-order, then a KB structure k knows α just in case α is true (in the first-order sense defined earlier) on every pair $\langle s, v\rangle$ that is an element of k. As with $\mathcal{L}$, we define a sentence to be *valid* if it is true on every s, v and k. A set of sentences is *satisfiable* if there is an s, v and a k on which every element of the set is true.

Since our main objective is to use $\mathcal{KL}$ as an interface language, it is the semantics of $\mathcal{KL}$ and how it leads to a definition of TELL and ASK that concerns us. If we were interested in formulating theories and reasoning about KB's, then the proof theory of $\mathcal{KL}$ would be useful to study. In fact, we are not so much interested in *describing* and reasoning about a KB here, but in actually *specifying* one and using it to reason about the world. For the sake of completeness, however, we present an axiomatization of $\mathcal{KL}$ that has the same rules of inference as $\mathcal{L}$ and the following axioms.[17]

Axioms. *Those of* $\mathcal{L}$ *(with modification to* AS*) and*

(KAX) $\boldsymbol{K}\alpha$ *where* α *is an axiom of* $\mathcal{L}$.

(KMP) $(\boldsymbol{K}\alpha \wedge \boldsymbol{K}(\alpha \supset \beta)) \supset \boldsymbol{K}\beta$.

(KUG) $\forall x\boldsymbol{K}\alpha \supset \boldsymbol{K}\forall x\alpha$.

(KCL) $\alpha \equiv \boldsymbol{K}\alpha$ *provided* α *is pure*.

The reason the Axiom of Specialization must be modified has to do with its interaction with the $\boldsymbol{K}$-operator. Consider, for example, a KB that thinks it knows all the major cities of Ohio. For this KB, the sentence

$$\forall \text{city}[\boldsymbol{K}\text{MajorCity}(\text{city}, \text{Ohio}) \vee \boldsymbol{K}\neg\text{MajorCity}(\text{city}, \text{Ohio})]$$

is true, in that for each city, the KB either believes that it is or believes that it is not a major city of Ohio. Indeed, this sentence follows from

[17] In [7] there is a Henkin style proof that this axiomatization is both sound and complete with respect to the semantics given above.

$\boldsymbol{K}\forall x$[MajorCity(x, Ohio) $\supset \boldsymbol{K}$MajorCity(x, Ohio)], the sentence that states that the KB thinks it knows all the major cities of Ohio. What we do *not* want to conclude from the fact that the KB has an opinion about every city is that the sentence

$$\boldsymbol{K}\text{MajorCity(FavoriteCity(Joe), Ohio)} \vee \boldsymbol{K}\neg\text{MajorCity(FavoriteCity(Joe), Ohio)}$$

is true. This sentence claims that the KB has an opinion about the favorite city of Joe, which does not follow at all, given that the KB may have no idea about what city is being referred to. Of course, a KB *could* believe that Joe's favorite city is not a major city of Ohio without having identified it. For example, it might know that Joe only likes California cities. The point here is that a KB *need not* have an opinion either way. The only time we really want to conclude that

$$\boldsymbol{K}\text{MajorCity}(t, \text{Ohio}) \vee \boldsymbol{K}\neg\text{MajorCity}(t, \text{Ohio})$$

from the above is when the KB knows the referent of the term t (i.e., has identified the equivalence class to which it belongs).

The problem is that the sentence about Joe's favorite city is a direct consequence of the first one by the Axiom of Specialization. To remedy this, we have to replace the axiom (and its occurrence within Axiom KAX) by a special version (called AS*) that has an attached proviso: if $\forall x\alpha$ is true, then α_t^x is also true *provided no function symbol of t gets placed within the scope of a* $\boldsymbol{K}$ *in the substitution.* The motivation behind this proviso is as follows. The language $\mathscr{L}$ has the property that if two terms are equal, then they can be used interchangeably without affecting the truth value of any sentence. The language $\mathscr{KL}$, on the other hand, does not have this property. A sentence α_k^x may be true for every parameter k and yet be false for some term t (even though t has to be equal to one of the parameters). In particular, the sentence

$$(t_1 = t_2) \supset \boldsymbol{K}(t_1 = t_2)$$

is not a valid sentence of $\mathscr{KL}$ for every pair of terms. However, the sentence

$$\forall x \forall y (x = y) \supset \boldsymbol{K}(x = y)$$

is valid (and a theorem of $\mathscr{KL}$). So a KB characterized by the language $\mathscr{KL}$ has the interesting property that identicals are indiscernable for it (if two things are identical, the KB cannot help but know this) and yet there are identity statements that are neither known to be true nor known to be false.

As for the other axioms of $\mathscr{KL}$, it is worth noticing that the KCL axiom

corresponds to the Assumption of Closure (since it states that a pure sentence is known exactly when it is true) and that KAX, KMP and KUG together correspond to the Assumption of Competence. KAX says that the axioms of $\mathscr{L}$ are known while KMP and KUG say that the KB can perform Modus Ponens and Universal Generalization based on what it knows. If we were ever to consider a less competent version of a KB, these axioms would be the ones to modify.

The robustness of our approach is suggested by all the desirable properties of $\mathscr{KL}$ that are not explicit in its formalization. For example, the assumption that a KB need not have accurate knowledge of reality is reflected in the fact that the sentence

$$(\boldsymbol{K}\alpha \supset \alpha)$$

is not a theorem of $\mathscr{KL}$ unless α is pure, unlike most modal logics. For example, unlike the formalizations of knowledge in [10, 11], the set $\{\neg\phi(c), \boldsymbol{K}\phi(c)\}$ can be satisfied given a world and a KB that happens to be mistaken about that world. The sentence

$$\boldsymbol{K}(\boldsymbol{K}\alpha \supset \alpha),$$

however, *is* indeed a theorem, meaning that although a KB need not have accurate knowledge, it will always believe that it does. A KB that did not have this belief would have no way of deciding if what it was told by a user was indeed true. But the kind of KB we are considering is *committed* to its knowledge, since it always knows that what it knows is true.

One possible source of concern in our definition of $\mathscr{KL}$ is that by virtue of KCL and the fact that KCL is itself pure, the language may be *reducible*, in the sense that for any sentence using nested $\boldsymbol{K}$ operators, there would be another logically equivalent one that had no nested operators. The effect of this would be to say that meta-knowledge was superfluous in $\mathscr{KL}$, that is, that any sentence that spoke about the the meta-knowledge of a KB could be rephrased in terms of world knowledge. If that were the case, our concept of meta-knowledge would be somewhat vacuous and a lot of the machinery used in the definition of $\mathscr{KL}$ would not be necessary. However, despite the presence of KCL, it can be shown (by an argument similar to one appearing in [7]) that the language $\mathscr{KL}$ is *not* reducible in this sense. Intuitively, it should be clear that the sentence

$$\boldsymbol{K}[\exists x(\phi(x) \wedge \neg\boldsymbol{K}\phi(x))]$$

is not equivalent to any sentence without nested $\boldsymbol{K}$-operators. To prove this formally involves finding two KB structures on which the sentence has different

truth values and yet which know exactly the same set of sentences of $\mathscr{L}$. It follows that any sentence without nested $\mathbf{K}$-operators will have the same truth value on both KB structures and so cannot be equivalent to the above one.

On the other hand, it is not as if reducibility was an implausible notion since there is also a proof in [7] that a *propositional* version of $\mathscr{KL}$ is indeed reducible. So, in the propositional case, the concept of meta-knowledge is trivial and dispensable. As a result of this simplification, we could limit ourselves to a language where no $\mathbf{K}$-operator occurs in the scope of another. Moreover, in this case, we could interpret a sentence $\mathbf{K}\alpha$ as if it said that α was logically implied (in the standard propositional sense) by the KB. This simplifying interpretation is not possible for the full quantificational language, however.

2.2.5. *The definition of* TELL *and* ASK

Having considered the semantics of $\mathscr{KL}$, we can now define the two operations TELL and ASK. The functionality of the two operations will be

$$\begin{aligned} &\text{TELL: } \mathrm{KB} \times \mathscr{KL} \to \mathrm{KB}\ ; \\ &\text{ASK: } \mathrm{KB} \times \mathscr{KL} \to \{\text{yes, no, unknown}\}\ . \end{aligned}$$

The first argument to both operations is to be understood as an *abstract knowledge base*, that is, a partial picture of a world, specified by a set of world structures. The second argument to TELL is an *assertion*; the second argument to ASK is a *yes–no question.*

Given our semantic specification of $\mathscr{KL}$, we can define ASK easily since a KB should answer *yes* to a sentence of $\mathscr{KL}$ (used as a question) exactly when it knows that the sentence is true. More formally, we have the following.

Definition. ASK:

$$\mathrm{ASK}[\![k, \alpha]\!] = \begin{cases} \text{yes}, & \text{if } \mathbf{K}\alpha \text{ is true on } k; \\ \text{no}, & \text{if } \mathbf{K}\neg\alpha \text{ is true on } k; \\ \text{unknown}, & \text{otherwise}. \end{cases}$$

Notice that a negative answer, indicates that the KB thinks that the sentence is false, not that it does not know (in which case it would have answered *unknown*). Moreover, by the Assumption of Closure, no pure question will be answered *unknown*. In fact, a question will be answered *unknown* only if there is an element of k on which the assertion is true and another on which it is false. For example, if we let k_0 denote the set of all world structures, then k_0 specifies the KB with the least amount of world knowledge since any sentence

of $\mathscr{L}$ is *unknown* except the valid ones and their negations. On the other hand, there are sentences of $\mathscr{KL}$ other than the pure or valid ones for which k_0 will answer *yes*. The sentence

$$\exists x\phi(x) \supset \exists x(\phi(x) \wedge \neg \mathbf{K}\phi(x)),$$

for instance, is known to be true by k_0. This sentence says that if there is a ϕ, then there must be a ϕ that is not a known ϕ (since currently, there are none).[18] The set of all sentences α such that ASK$[\![k_0, \alpha]\!]$ is *yes* is *undecidable*. In fact, it is not even semi-decidable. The easiest way to see this is to notice that for σ in $\mathscr{L}$, $\neg \mathbf{K} \neg \sigma$ is in the set iff σ is satisfiable and that the set of satisfiable sentences of $\mathscr{L}$ is not recursively enumerable. So there can be no effective procedure that handles yes–no questions in all cases. The above definition of ASK presumes that any implementation will have a heuristic component.

The definition of TELL is more complicated, as far as sentences containing $\mathbf{K}$ are concerned. For sentences of $\mathscr{L}$, however, we want the result of telling a KB that α is true to be a KB that knows α and everything else it knew about the world, but no more than that. In other words, the resulting KB should only contain world structures $\langle s, v\rangle$ on which α is true (in the first-order sense defined earlier). This suggests that for sentences of $\mathscr{L}$, we have that

$$\text{TELL}[\![k, \alpha]\!] = k \cap \{\langle s, v\rangle \mid \alpha \text{ is true on } s \text{ and } v\}.$$

The result of telling a KB specified by k that α is true is described by the largest set of world structures (has least amount of world knowledge) that is both a subset of k (knows everything it did before) and a subset of the world structures satisfying α (knows that α is true).

As discussed above, the motivation behind using $\mathscr{KL}$ as an assertion language is to be able to tell the KB sentences like

$$\exists x(\phi(x) \wedge \neg \mathbf{K}\phi(x)).$$

The interpretation of this assertion is not that there is a ϕ that will *never* be known, but that there is a ϕ that is not *currently* known. So within the context of a TELL, instances of $\mathbf{K}\alpha[x_1, \ldots, x_n]$ should be understood as the α's that were known *just prior to the* TELL *operation*. With this understanding, we can define TELL for all sentences of $\mathscr{KL}$ as follows:

[18]The fact that this sentence is known by the KB with the least amount of world knowledge is not really a property of the *logic* of $\mathscr{KL}$. From the point of view of the semantics, there is nothing special about this sentence other than it being satisfiable. It is the behavior of ASK with respect to specific KB structures that leads to this observation.

Definition. TELL:

$$\text{TELL}[\![k, \alpha]\!] = k \cap \{\langle s, v\rangle \mid \alpha \text{ is true on } s, v \text{ and } k\}.$$

This definition specifies that the result of a TELL on a KB described by k is just those elements of k where α is true, understanding any reference in α to what is known to mean known by the KB specified by k. This has the effect, as desired, that any pure assertion either produces k itself (when it is redundant) or the empty set (when it is contradictory). In general, the only assertions that are not either redundant or contradictory are those for which ASK would return *unknown* and, by the Assumption of Closure, no pure assertion is *unknown*.

One thing to notice about this definition of TELL and ASK is that the world knowledge of a KB is *monotonic* but that in general, knowledge is *non-monotonic*. As knowledge is acquired, we are assuming an ever more refined picture of the world in that any known sentence of $\mathscr{L}$ remains known.[19] But this is not true of $\mathscr{KL}$; not every known sentence of $\mathscr{KL}$ will remain known as knowledge is acquired. For example, as pointed out above,

$$\text{ASK}[\![k_0, [\exists x \phi(x) \supset \exists x(\phi(x) \wedge \neg \mathbf{K}\phi(x))]]\!] = \text{yes}$$

and yet

$$\text{ASK}[\![\text{TELL}[\![k_0, \phi(1)]\!], [\exists x \phi(x) \supset \exists x(\phi(x) \wedge \neg \mathbf{K}\phi(x))]]\!] = \text{unknown}.$$

In other words, after telling k_0 that *1* is a ϕ, the KB no longer believes that if there is a ϕ, then there has to be one it does not know about. Furthermore,

$$\text{ASK}[\![\text{TELL}[\![k_0, \forall x[\phi(x) \equiv (x = 1)]]\!], [\exists x \phi(x) \supset \exists x(\phi(x) \wedge \neg \mathbf{K}\phi(x))]]\!] = \text{no},$$

so that after telling k_0 that *1* is the *only* ϕ, it now believes that the original sentence is false. So while answers to questions in $\mathscr{L}$ can only move from *unknown* to *yes* or *no* as knowledge is acquired, answers to questions in $\mathscr{KL}$ can change arbitrarily as knowledge is acquired. Notice, however, that this non-monotonicity is not a property of the logic (proof-theory or semantics) of $\mathscr{KL}$, but of its *pragmatics*, which is perhaps as it should be [12].

[19]Essentially what this says is that TELL is not the appropriate operation for *belief revision* in the sense of tracking down a belief to be discarded in the presence of contradictory information. This is a thorny problem given the expressive power of $\mathscr{L}$. For example, if a KB is told $(\alpha \vee \beta)$, then $\neg\alpha$ and then $\neg\beta$, it is far from clear how its beliefs should be revised. The position here is that the competence of a KB should at the very least be able to detect the inconsistency as the first step towards making an intelligent decision about how to handle the conflict.

3. The Symbol Level

In the previous section, we considered an extremely abstract notion of a KB, specified by a set of world structures. We also examined how the TELL and ASK operations could be defined over such sets in a way that made no assumptions about how the knowledge was represented. A number of questions remain unanswered, however. First of all, exactly what sets of world structures do indeed specify a KB? Does every distinct set of worlds constitute a distinct body of world knowledge? Can this world knowledge always be represented symbolically? Given a symbolic representation of the knowledge in a KB, how can the TELL and ASK operations be defined over these representations. These are the questions that will be answered in this section. So the purpose of the section is to solidify the notion of a KB and show that, despite the presence of $\mathcal{KL}$ as an assertion language, the kind of KB we are considering is one that can be represented in first-order terms.

3.1. The representation of knowledge

In this section, we examine closely sets of world structures and determine which of these can be said to specify knowledge bases. In so doing we will isolate two special categories of KB structures: the *representable* and the *ω-representable* sets of world structures. The problem we have to face is that only the first kind can be represented symbolically, but the language $\mathcal{KL}$ forces us to also deal with the second.

3.1.1. *Representable KB structures*

In the preceding section, we defined TELL and ASK at the knowledge level independently of any symbolic representation of knowledge. However, we made no special provisions to ensure that the kind of knowledge we were considering was indeed representable in first-order terms at the symbol level. The kind of representation we have in mind here, obviously, is using a sentence (or a finite set of sentences) of $\mathcal{L}$ to refer to a set of worlds, that is, using a finite collection of symbols (instead of a possibly uncountably infinite set of world structures) to represent what is known. Thus, a sentence of $\mathcal{L}$ *represents* those worlds that satisfy the sentence. We will call a KB structure k *representable* iff there is a sentence of $\mathcal{L}$ such that k is the set of world structures satisfying that sentence. Clearly, our first concern in $\mathcal{KL}$ must be with a semantics ranging over representable KB structures since these are those for which the KB can be realized. However, as we will show below, we must be careful in restricting our attention to just these sets.

The first thing to notice is that not every KB structure is representable. Consider, for example, an arbitrary ordering of the parameters i_1, i_2, and so on.

Now let

$$\text{ODD} = \{i_1, i_3, \ldots\}, \quad \text{every second element of the ordering .}$$

Finally, assume that ϕ is some monadic predicate and consider the set

$$\{\langle s, v\rangle \mid \phi(i) \text{ is true on } s \text{ and } v \text{ iff } i \in \text{ODD}\}$$

This set of world structures is for a KB that knows that the ϕ are the same as the ODD-parameters. Because no single sentence of $\mathscr{L}$ can represent that knowledge about ϕ, the KB is not representable.

Now consider the following sentence of $\mathscr{KL}$:

$$\begin{aligned}
&\forall x \neg\theta(x, x) \\
&\quad \wedge \forall x \forall y \forall z[\theta(x, y) \wedge \theta(y, z) \supset \theta(x, z)] \\
&\quad \wedge \boldsymbol{K}\phi(1) \wedge \boldsymbol{K} \neg\phi(2) \\
&\quad \wedge \forall x[\boldsymbol{K}\phi(x) \supset \exists y\theta(x, y) \wedge \boldsymbol{K}\phi(y)] \\
&\quad \wedge \forall x[\boldsymbol{K} \neg\phi(x) \supset \exists y\theta(x, y) \wedge \boldsymbol{K} \neg\phi(y)] .
\end{aligned}$$

This sentence first says that θ is irreflexive and transitive (think of it as 'less than' where parameters are treated as numbers). Next, it says that there is a known ϕ and a known non-ϕ. Finally, it says that for each known ϕ there is a 'larger' one and similarly for the known non-ϕ. So the sentence implies that the known ϕ and the known non-ϕ are both infinite in number. This sentence is satisfiable since it is true in a world where θ is interpreted as the *less-than* relation and the KB is the one that knows about the ODD-parameters, above. However, the crucial point is that this sentence is not satisfied by *any* representable KB structure.

This leaves us in somewhat of a quandary. The representable sets are defined to be the *only* KB structures that can be represented at the symbol level. The language $\mathscr{KL}$, on the other hand, forces us to consider non-representable sets. If we imagine the semantics of $\mathscr{KL}$ adjusted so that only representable KB structures were used, the negation of the above sentence would have to be valid (and a theorem). In other words, either we change the semantics (and proof theory) of $\mathscr{KL}$ or we admit that there are knowledge bases that cannot be represented symbolically.

It is very difficult to imagine how the proof theory of $\mathscr{KL}$ could be modified to make sentences like the negation of the above come out as theorems. One can even imagine proving that no *axiomatic* theory would be up to the task.[20]

[20]As will become clear below, a theory that allows all representable KB structures without admitting any non-representable ones is like a number theory that says that numbers can be arbitrarily large and yet somehow rules out infinitely large ones. It seems beyond the power of recursively axiomatizable theories in $\mathscr{L}$ or $\mathscr{KL}$ to capture such distinctions.

So we are almost forced into looking beyond representable sets to see exactly what sort of KB we can, in fact, describe using sentences of $\mathcal{KL}$.

3.1.2. *ω-representable KB structures*

The non-representable KB structure above can be thought of as specifying one that knows

$$\{\phi(i_1), \neg\phi(i_2), \phi(i_3), \neg\phi(i_4), \ldots\}.$$

This suggests a generalization of representable sets that allows the knowledge to be represented using a potentially *infinite* set of sentences of $\mathcal{L}$. We will call a KB structure k *ω-representable* iff there is a set of sentences of $\mathcal{L}$ such that k is the set of world structures satisfying the set. Obviously, every representable set is ω-representable but not vice versa.

The ω-representable KB structures turn out to be just the right generalization of representable ones. For example, we might say that two knowledge bases are *equivalent* iff they know exactly the same sentences of $\mathcal{L}$. Then, no two distinct ω-representable sets are equivalent. Moreover, it is easy to verify that for *any* KB structure, there is a unique ω-representable one that is equivalent to it. So we can partition the KB structures into equivalence classes (with respect to the above definition of equivalence) and find exactly one ω-representable KB structure in each class.

Consider, for example k_0, the set of all world structures, a ω-representable set represented by, among others, the empty set of sentences. It is shown in [7] that the set k_0 and the set formed by removing any single world structure from k_0 know exactly the same sentences of $\mathcal{L}$. In other words, the two KB structures are equivalent. The question is why we would ever want to consider there to be *two* sets here in the first place; semantically, we would like to be able to say that there is only a *single* abstract body of knowledge under consideration, and therefore, a single set of worlds. This can be done very simply by restricting our attention to ω-representable sets. In this way we never get two distinct KB structures unless there is a sentence of $\mathcal{L}$ that they disagree on. So ω-representable sets are exactly analogous to world structures: just as for any two distinct world structures there is a sentence of $\mathcal{L}$ that one says is *true* and the other not, for any two distinct ω-representable KB structures, there is a sentence of $\mathcal{L}$ that one says is *known* and the other not.

The key property of ω-representable sets, however, is that as far as $\mathcal{KL}$ is concerned, we can restrict our attention to them without any loss of generality. In [7], the following theorem is proven.

Maximal Set Theorem.[21] *If a sentence of $\mathcal{KL}$ is satisfiable at all, it is satisfied on some ω-representable KB structure.*

In other words, restricting our attention to ω-representable sets does not change the semantics of $\mathcal{KL}$. Exactly the same sentences are valid and satisfiable. In particular, any (finite) condition we can impose on a KB using a sentence of $\mathcal{KL}$ can be satisfied by one specified by a ω-representable set, if at all.[22] This is not a trivial result, however, in that given an arbitrary KB structure, the ω-representable set predicted by the theorem will not necessarily be equivalent to it (that is, have the same world knowledge). Rather, the set is the result of an elaborate construction where a KB has to be given world knowledge about an *infinite* number of predicates not appearing in the sentence.

Despite how naturally the ω-representable sets fit into the semantics of $\mathcal{KL}$, they do have a very serious drawback. Specifically, the TELL function does not necessarily map one ω-representable set into another. For instance, let

$$k = \{\langle s, v\rangle \mid \text{for every } i \in \text{ODD}, \phi(i) \text{ is true on } s \text{ and } v\}.$$

So, k specifies a KB that knows that every element of ODD has the property ϕ. It is ω-representable and is represented by

$$\{\phi(i) \mid i \in \text{ODD}\}.$$

If we now let

$$k' = \text{TELL}[\![k, [\exists x(\psi(x) \wedge \mathbf{K}\phi(x))]]\!],$$

then we have that k' is the result of telling k that there is a ψ among the known ϕ. Since the parameters that are known to be ϕ are the ODD ones, k has been told that there is a ψ among the ODD-parameters. In other words, k' is the set

$$\{\langle s, v\rangle \mid \text{for every } i \in \text{ODD}, \phi(i) \text{ is true on } s \text{ and } v \text{ and for some } i \in \text{ODD}, \psi(i) \text{ is true on } s \text{ and } v\}.$$

[21] In [7], ω-representable sets are called 'maximal' and are defined as follows: a KB structure k is *maximal* iff it is the set of all $\langle s, v\rangle$ such that $(\mathbf{K}\alpha \supset \alpha)$ is true on s, v and k, for every α in $\mathcal{L}$. This definition can be shown to be equivalent to the current one.

[22] There are, however, *infinite* sets of sentences that are satisfiable but not by any ω-representable KB structure. One such is the set of all pure sentences that are true on k', described below. If the expressive power of a language is taken to be a function of what distinctions *sentences* can make, the move to ω-representable sets is without loss of generality; but if it is a function of what *sets of sentences* can do, the restriction is real.

Thus, k' believes that every ODD-parameter is ϕ and that at least one is ψ.

The KB structure k' is anomalous, however. First of all, it is not ω-representable. It can be shown that the ω-representable set that is equivalent to it (call it k'_*) contains a world structure where ψ is true for no ODD-parameter. In other words, k'_* no longer believes that there is an ODD ψ as k' did. And yet, k' and k'_* are equivalent since they agree completely on every sentence of $\mathscr{L}$. Where they disagree is on the sentence

$$\exists x(\psi(x) \wedge \boldsymbol{K}\phi(x)).$$

The KB described by k' believes that it is true while the one described by k'_* does not.

It might be thought that the restriction to ω-representable sets was perhaps too facile, that k' and k'_* in fact specify knowledge bases that must be considered distinct. Perhaps all of $\mathscr{KL}$ must be used to represent world knowledge (and to decide whether two KBs have identical world knowledge). This would allow k' and k'_* to be distinguished in terms of the sentence of $\mathscr{KL}$ on which they disagree.

As it turns out, however, it is k' itself that is at fault here and a generalization that allows all of $\mathscr{KL}$ to be used to represent knowledge would not solve the problem. To see this, consider

$$k'' = \text{TELL}[\![k', [\forall x\phi(x)]]\!].$$

This KB knows that every parameter (not just the ODD ones) is a ϕ. Assuming that its knowledge about the ψ is unchanged by this assertion, k'' cannot just believe that there is a ψ among the known ϕ since it would then lose all connection with the ODD-parameters. But what exactly does the KB described by k'' believe about ψ? This KB is truly anomalous since we can no longer represent its belief *even using an infinite set of $\mathscr{KL}$ sentences.*[23] In some sense, it was coincidental that we could represent the knowledge of k': we characterized what it knew about ψ using what it knew about ϕ. But even if we consider representing the knowledge of k' using sentences of $\mathscr{KL}$, we will still run into situations (like k'') where even $\mathscr{KL}$ is no help. The conclusion: k' is anomalous and, therefore, not every ω-representable KB structure can be used as an argument to TELL.

[23]The best we might do is use an infinitely long disjunction $[\psi(i_1) \vee \psi(i_3) \vee \cdots]$ or somehow time stamp a KB to be able to refer to the ϕ that were known just prior to the assertion.

So where we have arrived in our examination of representable and ω-representable sets is the following.

(1) The language $\mathcal{KL}$ forces us into considering knowledge bases that are not representable.

(2) The KB's described by ω-representable KB structures are both necessary and sufficient from the point of view of $\mathcal{KL}$, but are not closed under the TELL operation.

So what we have to do[24] is try to understand ω-representable sets in such a way that the knowledge bases they specify cannot be used as arguments to TELL. One account that does the trick is to understand these sets as limiting cases of representable ones.

Let $k_1, k_2, \ldots$ be an infinite sequence of representable KB structures where each element in the sequence (as a set of world structures) is a subset of the previous one (that is, a monotone decreasing sequence). We call a KB structure k_ω the *limit of the sequence* if it is the largest set that is a subset of every element in the sequence (that is, it is the greatest lower bound). It is easy to show that the ω-representable sets are precisely the limits of descending sequences of representable sets. This kind of sequence is, in fact, a path of *knowledge acquisition* since each element in the sequence must know at least as much as its predecessors. So a representable set specifies a KB that is located on a knowledge acquisition path while a ω-representable set specifies the ultimate destination of such a path, which may never be attained. For example, if k_ω is the ω-representable set represented by some set of sentences $\{\alpha_i\}$, then it is also the limit of a sequence of representable sets

$$
\begin{aligned}
k_0 &= \text{the set of all world structures,}\\
k_1 &= \text{TELL}[\![k_0, \alpha_1]\!],\\
k_2 &= \text{TELL}[\![k_1, \alpha_2]\!],\\
k_3 &= \text{TELL}[\![k_2, \alpha_3]\!],\\
&\vdots
\end{aligned}
$$

In other words, to be the limit of an infinite descending sequence of representable sets is also to be the limit of an infinite set of TELL operations. So the reason a KB described by a non-representable set cannot be told anything is that *you never really get there* except in the limit. In particular, although it makes sense to consider the limit of a sequence of operations, it does not make sense to *arrive* at the limit point and perform yet another operation. Although the language $\mathcal{KL}$ forces us to admit the existence of these limit points, we can understand TELL and ASK as dealing only with the representable ones.

[24] Another option we have not considered is to try a notion of representability somewhere between the representable and ω-representable sets. For example, we might consider a KB that can be represented by a ***recursive*** or ***recursively enumerable*** set of sentences, although it is unlikely that we could get by without modifying the semantics for $\mathcal{KL}$.

The only thing left to do to with representable sets is to establish that (unlike ω-representable sets) they are closed under the TELL operation. If the only assertions under consideration were sentences of $\mathscr{L}$, this would be easy to show since the result of telling a KB k that α is true (where α is first-order) could be represented by $(k \wedge \alpha)$. The fact that we need only consider representable KB structures *even when assertions are taken from* $\mathscr{KL}$ is the major result of this research, which we now describe.

3.2. The representation theorem

It is perhaps worthwhile after the excursion into the mathematical properties of $\mathscr{KL}$, TELL and ASK and KB structures, to reconsider what has been accomplished and where we stand.

We started with a notion of a KB as a first-order sentence and Tell- and Ask-operations that were purely proof theoretic: Tell simply conjoined the assertion to the KB and Ask used an oracle for first-order theoremhood to determine its answer. We then replaced the symbolic notion of a KB by a functional (representation independent) one: a KB was thought of as anything having a certain 'world knowledge', understood as a certain behavior under the interaction operations. We showed how a first-order interaction language was too weak and subsequently replaced it with a more expressive one, $\mathscr{KL}$. In the process of defining the TELL and ASK operations in terms of the semantics of $\mathscr{KL}$, we were lead to an abstract picture of a KB as anything embodying a partial picture of a world, and therefore specifiable by a set of world structures. We showed that without loss of generality we could restrict our attention to the ω-representable sets of world structures and that a subset of these, the representable sets, specified knowledge bases that could be represented using sentences of $\mathscr{L}$.

The question that now remains is the following: if we chose to represent a KB using a sentence of $\mathscr{L}$, what can be said about the TELL and ASK operations when $\mathscr{KL}$ is the language of interaction? Specifically:

(1) Is the KB after a TELL operation always representable and how might a representation be constructed from the representation before the TELL?

(2) How is the ASK operation when using a sentence of $\mathscr{KL}$ related to the ASK operation over first-order sentences?

The answer here is somewhat surprising (and will be examined in detail): despite the increase in expressive power given by $\mathscr{KL}$,

(1) The result of a TELL can always be represented by conjoining some first-order sentence to the KB.

(2) Given a first order KB, the ASK operation can be performed using only the oracle for first-order theoremhood.

So even though we are communicating with a KB in an intensional language that is more powerful than $\mathscr{L}$, this communication can be understood and

represented completely in first-order terms. So for example, one could imagine using a standard (first-order) resolution theorem prover to implement the ASK operation.[25] A version of TELL and ASK that uses this oracle will be defined below with the understanding that the definition is not intended to demonstrate an *efficient* algorithm, only that an algorithm is *possible*, given the oracle.

3.2.1. *Method*

The main observation leading to the Representation Theorem is that given a first-order sentence k that represents some world knowledge, it is possible to *reduce* any formula α of $\mathcal{KL}$ to a first-order formula which we will call $|\alpha|_k$, without changing its assertional force. If α has n free variables, then the result of telling the KB that (or asking the KB if) $|\alpha|_k[i_1, \ldots, i_n]$ is true is the same as if $\alpha[i_1, \ldots, i_n]$ had been used. Notice that $|\alpha|_k$ is indexed by the KB k and so the formula will change as the KB changes. In fact, one of the properties of $\mathcal{KL}$ is that it is impossible to find a first order sentence that is equivalent to α for *every* KB.

As it turns out, being able to handle the general case of $|\alpha|_k$ depends only on a correct treatment of $|\boldsymbol{K}\alpha|_k$ when α is a formula of $\mathcal{L}$. To see an example of how this will work, imagine that k is the sentence

City(LosAngeles)

where, for this example, constants are treated like parameters (that is, Los Angeles is the only known city). Now suppose we want a first-order equivalent to $\boldsymbol{K}\text{City}(x)$ for this KB. That is to say, we want a formula of $\mathcal{L}$, $|\boldsymbol{K}\text{City}(x)|_k$, with one free variable x, such that for any parameter i,

$$(|\boldsymbol{K}\text{City}(x)|_k)^x_i \text{ is true} \quad \text{iff} \quad \boldsymbol{K}\text{City}(i) \text{ is true}.$$

But this sentence is true for the KB exactly when i is the LosAngeles parameter. This suggests that $|\boldsymbol{K}\text{City}(x)|_k$ can be the formula $(x = \text{LosAngeles})$, instances of which are *valid* when x is LosAngeles and *unsatisfiable* otherwise. On the other hand, had k been

$$\text{City(LosAngeles)} \wedge \text{City(NewYorkCity)} \wedge [\text{City(Boston)} \vee \text{City(Detroit)}],$$

the formula would have been $[(x = \text{LosAngeles}) \vee (x = \text{NewYorkCity})]$ since neither Boston nor Detroit are known instances of City (although it is known

[25]This assumes that we could define a theorem prover to work with $\mathcal{L}$, a slightly non-standard first-order dialect. Indeed, the treatment of *equality* in $\mathcal{L}$ is critical to this approach.

that one of them is a city). Similarly, $|\boldsymbol{K}\text{City}(z)|_k$ when k is $\forall x\text{City}(x)$ is $(z = z)$ since every parameter is a known city in this case.

To simplify our notation, we will let RES$[\![k, \alpha]\!]$ stand for $|\boldsymbol{K}\alpha|_k$ when α is a formula of $\mathscr{L}$. So, for example, RES$[\![$City(LosAngeles), City$(z)]\!]$ is $(z =$ LosAngeles). The situation becomes more interesting when the formula we want to reduce contains multiple free variables. For example, if k is the sentence $[\psi(1, 7) \wedge \psi(5, 2)]$, then RES$[\![k, \psi(5, y)]\!]$ is $(y = 2)$ while RES$[\![k, \psi(x, y)]\!]$ is

$$(x = 1 \wedge y = 7) \vee (x = 5 \wedge y = 2)\,.$$

It may not always be possible just to 'read' off the correct values for the free variables. For example,

$$\text{RES}[\![\forall x \forall y[x = y \equiv \neg\psi(y, x)], \psi(x, 4)]\!] = (x \neq 4)$$

and

$$\text{RES}[\![(\psi(1, 6) \wedge \forall x \forall y[[(x = y) \vee \psi(x, y)] \supset \psi(y, x)]), \psi(6, y)]\!] = [(y = 1) \vee (y = 6)]\,.$$

In general, calculating RES$[\![k, \alpha]\!]$ may require arbitrary amounts of deduction and there will be no effective procedure for calculating it.

Before proceeding to formally define a RES$[\![k, \alpha]\!]$ and proving that it works for any KB, we will simply assume that it works and show how it leads to the Representation Theorem. First of all, we can define the general case of $|\alpha|_k$ in terms of RES$[\![k, \alpha]\!]$ as follows:

$$\begin{aligned}
&|\alpha|_k = \alpha \quad \text{when } \alpha \in \mathscr{L},\\
&|\neg\alpha|_k = \neg|\alpha|_k,\\
&|\alpha \supset \beta|_k = (|\alpha|_k \supset |\beta|_k),\\
&|(\forall x)\alpha|_k = (\forall x)|\alpha|_k,\\
&|\boldsymbol{K}\alpha|_k = \text{RES}[\![k, |\alpha|_k]\!].
\end{aligned}$$

This gives us $|\alpha|_k$ for any α in $\mathscr{KL}$ in terms of how it handles $\boldsymbol{K}\alpha$ when α is first-order. For example,

$$|\forall x(\text{City}(x) \supset \boldsymbol{K}\text{City}(x))|_{\text{City(LosAngeles)}} = \forall x(\text{City}(x) \supset x = \text{LosAngeles})\,.$$

What this says is that if all that is known is that Los Angeles is a city, then the sentence that says that every city is known reduces to one that says that every city is Los Angeles (that is, that it is the only city).

To establish more rigorously the fact that $|\alpha|_k$ correctly captures α given that k represents everything that is known, we have to be clear about what knowledge a first-order sentence k represents. So, we define

$$\mathscr{R}[k] = \{\langle s, v\rangle \mid k \text{ is true on } s \text{ and } v\}\,.$$

A first-order sentence k presents a picture of a world specified by those world structures on which the sentence is true. In other words, a sentence of $\mathcal{L}$ is taken to represent the knowledge that the world satisfies it. The first observation about $\mathcal{R}[k]$ is the following.

Lemma 3.1. *For any sentences α and k of $\mathcal{L}$, $\mathbf{K}\alpha$ is true on $\mathcal{R}[k]$ iff $(k \supset \alpha)$ is a theorem of $\mathcal{L}$.*

Proof. The sentence $\mathbf{K}\alpha$ is true on $\mathcal{R}[k]$ when α is true on every element of $\mathcal{R}[k]$, that is, when every world structure satisfying k also satisfies α. Then, assuming the proof theory for $\mathcal{L}$ is complete, this happens exactly when $(k \supset \alpha)$ is a theorem of $\mathcal{L}$.

This lemma says that a first-order sentence is known exactly when it is a (first-order) consequence of the sentence used to represent the knowledge. This does not mean that we can define $\mathrm{RES}[\![k, \alpha]\!]$ to be $(k \supset \alpha)$ since this sentence will not, in general, be *false* just because $\mathbf{K}\alpha$ is false. The above lemma only guarantees that, in this case, it will not be *valid*. However, what we will prove below (in Lemma 3.6) is that for any formula α of $\mathcal{L}$,

$$\mathbf{K}\alpha[i_1, \ldots, i_n] \text{ is true on } \mathcal{R}[k]$$
$$\text{iff} \quad \mathrm{RES}[\![k, \alpha]\!][i_1, \ldots, i_n] \text{ is true (on } \mathcal{R}[k]).$$

Given this property, and our definition of $|\alpha|_k$ in terms of $\mathrm{RES}[\![k, \alpha]\!]$, we can prove the following.

Lemma 3.2. *For any formula α of $\mathcal{KL}$, any k in $\mathcal{L}$, any set of parameters $i_1, \ldots, i_n$, and any world structure $\langle s, v\rangle$, $\alpha[i_1, \ldots, i_n]$ is true on s, v, and $\mathcal{R}[k]$ iff $|\alpha|_k[i_1, \ldots, i_n]$ is true on s and v.*

Proof. By induction on the structure of α. If α is a formula of $\mathcal{L}$, the lemma trivially holds, since $|\alpha|_k$ is α. Otherwise, assume the lemma holds for two arbitrary formulas α and β. It will then clearly hold for $\neg\alpha$ and $(\alpha \supset \beta)$. As for universal generalization, $\forall x\alpha[i_1, \ldots, i_n, x]$ is true iff for every parameter i, $\alpha[i_1, \ldots, i_n, i]$ is true iff (by induction) $|\alpha|_k[i_1, \ldots, i_n, i]$ is true iff $\forall x|\alpha|_k[i_1, \ldots, i_n, x]$ is true iff $|\forall x\alpha|_k[i_1, \ldots, i_n, x]$ is true. Finally, $\mathbf{K}\alpha[i_1, \ldots, i_n]$ is true on $\mathcal{R}[k]$ iff $\alpha[i_1, \ldots, i_n]$ is true on every world structure in $\mathcal{R}[k]$ iff (by induction) $|\alpha|_k[i_1, \ldots, i_n]$ is true on every world structure in $\mathcal{R}[k]$ iff $\mathbf{K}|\alpha|_k[i_1, \ldots, i_n]$ is true on $\mathcal{R}[k]$. But $|\alpha|_k$ is, by definition, a formula of $\mathcal{L}$. So, by Lemma 3.6 to be proven below, $\mathbf{K}|\alpha|_k[i_1, \ldots, i_n]$ is true on $\mathcal{R}[k]$ iff $\mathrm{RES}[\![k, |\alpha|_k]\!][i_1, \ldots, i_n]$ is true. Thus, $\mathbf{K}\alpha[i_1, \ldots, i_n]$ is true iff $|\mathbf{K}\alpha|_k[i_1, \ldots, i_n]$ is true.

Lemma 3.2 establishes (given an appropriate definition of RES$[\![k, \alpha]\!]$) that $|\alpha|_k$ correctly captures within a first-order formula what α was saying about the KB. We can now prove the Representation Theorem which shows how TELL and ASK can be defined given k as the representation of what is known.

Representation Theorem. *For any KB k and any sentence of $\mathcal{KL}$ α,*

$$\mathrm{TELL}[\![\mathscr{R}[k], \alpha]\!] = \mathscr{R}[k \wedge |\alpha|_k]$$

$$\mathrm{ASK}[\![\mathscr{R}[k], \alpha]\!] = \begin{cases} \text{yes}, & \text{if } \vdash (k \supset |\alpha|_k); \\ \text{no}, & \text{if } \vdash (k \supset \neg|\alpha|_k); \\ \text{unknown}, & \text{otherwise}. \end{cases}$$

Proof. For ASK$[\![\mathscr{R}[k], \alpha]\!]$, the answer will be *yes* precisely when $\mathbf{K}\alpha$ is true on $\mathscr{R}[k]$ which, by Lemma 3.2 coincides with $\mathbf{K}|\alpha|_k$ being true on $\mathscr{R}[k]$. But since $|\alpha|_k$ is first-order, by Lemma 3.1, this happens exactly when the sentence $(k \supset |\alpha|_k)$ is a theorem of $\mathscr{L}$. The situation is analogous when the answer from ASK is *no*. The proof for TELL, on the other hand, follows from Lemma 3.2 directly.

$$\begin{aligned} \mathrm{TELL}[\![\mathscr{R}[k], \alpha]\!] &= \mathscr{R}[k] \cap \{\langle s, v\rangle \mid \alpha \text{ is true on } s, v \text{ and } \mathscr{R}[k]\} \\ &= \mathscr{R}[k] \cap \{\langle s, v\rangle \mid |\alpha|_k \text{ is true on } s, v\} \\ &= \{\langle s, v\rangle \mid k \text{ and } |\alpha|_k \text{ are true on } s, v\} \\ &= \mathscr{R}[k \wedge |\alpha|_k]. \end{aligned}$$

The Representation Theorem not only guarantees that the result of a TELL will be first-order representable no matter what the argument, it also stipulates that it will be representable by an extension of the existing KB (that is, by conjoining something to it). Moreover, the Representation Theorem says that, no matter what the argument, the result of an ASK depends only on whether or not a first-order sentence is implied by the KB. However, the theorem assumes a very specific but yet to be proven property of RES$[\![k, \alpha]\!]$, which is that

$\mathbf{K}\alpha[i_1, \ldots, i_n]$ is true on $\mathscr{R}[k]$
iff RES$[\![k, \alpha]\!][i_1, \ldots, i_n]$ is true (on $\mathscr{R}[k]$).

Once the definition of RES$[\![k, \alpha]\!]$ is provided below and the critical Lemma 3.6 is proven, it will be clear that the Representation Theorem not only provides a proof theoretic view of the TELL and ASK operations, but a *first-order* proof theoretic view.

The reason this theorem is so important is that it forms a bridge not only between the knowledge level and symbol level but between $\mathcal{KL}$ and $\mathscr{L}$. It says that we can use $\mathcal{KL}$ to handle assertions like closed (or open) world assumptions and defaults (as we will discuss later), but still restrict our implementation

methods to standard first-order techniques (such as resolution theorem proving). Not that this makes TELL or ASK decidable in the general case; in fact, since both need to calculate $|\alpha|_k$ when α is not first order, both are undecidable. But the only oracle we need is one that can tell us if $(k \supset \alpha)$ is a first-order theorem (where both sentences are first-order). In other words, provided we use $\mathscr{L}$ as our symbolic representation language, this oracle is all we need to correctly implement a TELL and ASK operation for any sentence of $\mathscr{KL}$.

3.2.2. *Proof*

The definition of RES$[\![k, \alpha]\!]$ that we will use is by no means the shortest formula that will work, as it uses every parameter appearing in either in k or α. The definition is

RES$[\![k, \alpha]\!]$ = **If** α has no free variables
then if $\vdash (k \supset \alpha)$
then $\forall x(x = x)$
else $\forall x(x \neq x)$
else (Assume that x is free in α and that the parameters appearing in k or α are $i_1, \ldots, i_n$. Let i be the "least" parameter not in k or α.)

$$((x = i_1) \wedge \text{RES}[\![k, \alpha^x_{i_1}]\!]) \vee \cdots \vee ((x = i_n) \wedge \text{RES}[\![k, \alpha^x_{i_n}]\!])$$
$$\vee ((x \neq i_1) \wedge \cdots \wedge (x \neq i_n) \wedge \text{RES}[\![k, \alpha^x_i]\!]^i_x).$$

So, if α is a sentence, then RES$[\![k, \alpha]\!]$ is either valid or inconsistent depending on whether or not α is implied by k. Otherwise, assuming that x is a free variable of α, we call RES$[\![k, \alpha]\!]$ recursively with one fewer free variable by replacing x in α by some parameter. First, we use the parameters appearing in α or k, and then we use another parameter, i, that does not. The first thing to establish here is that the choice of i is irrelevant. To do so, we will show that a theorem of $\mathscr{L}$ remains a theorem when the parameters in it are renamed consistently.

Lemma 3.3. *Let $*$ be a bijection over parameters and, for any α in $\mathscr{L}$, let α^* be the result of replacing each i in α by i^*. Then, α is a theorem iff α^* is.*

Proof. By induction on the length of the proof of α. If α is an axiom of $\mathscr{L}$, then there are two cases: if α is not an instance of the equality schema, then the presence of parameters in α is incidental and *any* renaming of the parameters is still an axiom; if, on the other hand, α is of the form $(i = i) \wedge (i \neq j)$ for distinct parameters i and j, then α^* is $(i^* = i^*) \wedge (i^* \neq j^*)$ which is also an instance of the equality schema since $*$ is a bijection. Now suppose that α

follows from $(\sigma \supset \alpha)$ and σ by Modus Ponens. By induction, σ^* and $(\sigma^* \supset \alpha^*)$ must also be theorems, so α^* follows again by Modus Ponens. Finally, suppose $\forall x\alpha$ follows by Universal Generalization from sentences α_i^x, for i ranging over all the parameters occurring in α and at least one not present in α. By induction, the sentences $(\alpha_i^x)^*$, that is, $\alpha^{*\,x}_{i^*}$, must also be theorems. But since $*$ is a bijection, i appears in α iff i^* appears in α^*. So, the i^* must now range over all the parameters appearing in α^* and one that does not occur in α^*. Thus, by Universal Generalization, $\forall x\alpha^*$, that is, $(\forall x\alpha)^*$, is also a theorem.

Corollary 3.4. *If $*$ is a bijection and α is a formula such that α^* is α, then for any set of parameters, $\alpha[i_1, \ldots, i_n]$ is a theorem iff $\alpha[i_1^*, \ldots, i_n^*]$ is.*

Given this renaming property, we can show that $\mathrm{RES}[\![k, \alpha]\!]$ produces a formula whose instances are theorems exactly when instances of $(k \supset \alpha)$ are.

Lemma 3.5. *For any formula α and sentence k of $\mathscr{L}$, and any parameters $j_1, \ldots, j_m$, the sentence $\mathrm{RES}[\![k, \alpha]\!][j_1, \ldots, j_m]$ is a theorem iff $(k \supset \alpha)[j_1, \ldots, j_m]$ is a theorem.*

Proof. By induction on the number of free variables in α. If α is a sentence, then $\mathrm{RES}[\![k, \alpha]\!]$ is a theorem iff it is $\forall x(x = x)$ which happens only when $(k \supset \alpha)$ is a theorem. Otherwise, suppose α is $\alpha[x_1, \ldots, x_m, x]$. By definition, $\mathrm{RES}[\![k, \alpha]\!]$ is the formula

$$\begin{aligned}&((x = i_1) \wedge \mathrm{RES}[\![k, \alpha^x_{i_1}]\!])\\ &\quad \vee \cdots \vee ((x = i_n) \wedge \mathrm{RES}[\![k, \alpha^x_{i_n}]\!])\\ &\quad \vee ((x \neq i_1) \wedge \cdots \wedge (x \neq i_n) \wedge \mathrm{RES}[\![k, \alpha_i^x]\!]^i_x)\,.\end{aligned}$$

There are two cases to consider for $\mathrm{RES}[\![k, \alpha]\!][j_1, \ldots, j_m, j]$. First of all, assume that j is one of the parameters in k or α, say i_3. In this case, $\mathrm{RES}[\![k, \alpha]\!][j_1, \ldots, j_m, j]$ is a theorem iff $\mathrm{RES}[\![k, \alpha^x_{i_3}]\!][j_1, \ldots, j_m]$ is a theorem since all the other disjuncts of $\mathrm{RES}[\![k, \alpha]\!]$ will be inconsistent for this choice of x. However, by induction, $\mathrm{RES}[\![k, \alpha_j^x]\!][j_1, \ldots, j_m]$ is a theorem iff $(k \supset \alpha_j^x)[j_1, \ldots, j_m]$ is a theorem and this sentence is just $(k \supset \alpha)[j_1, \ldots, j_m, j]$. The other case to consider is when j is not a parameter of α or k. In this case, $\mathrm{RES}[\![k, \alpha]\!][j_1, \ldots, j_m, j]$ is a theorem iff $\mathrm{RES}[\![k, \alpha_i^x]\!]^i_x[j_1, \ldots, j_m, j]$ is a theorem since all other disjuncts of $\mathrm{RES}[\![k, \alpha]\!]$ will be inconsistent for this choice of x. Now let $*$ be the bijection over parameters that swaps i and j and leaves all other parameters unchanged. Then, by Corollary 3.4 $\mathrm{RES}[\![k, \alpha_i^x]\!]^i_x[j_1, \ldots, j_m, j]$ is a theorem iff $\mathrm{RES}[\![k, \alpha_i^x]\!]^i_x[j_1^*, \ldots, j_m^*, i]$ is a theorem. This sentence is just $\mathrm{RES}[\![k, \alpha_i^x]\!][j_1^*, \ldots, j_m^*]$ and, by induction, is a theorem exactly when $(k \supset \alpha_i^x)[j_1^*, \ldots, j_m^*]$ is. But this sentence is $(k \supset \alpha)[j_1^*, \ldots, j_m^*, i]$, which by Corollary 3.4 again, is a theorem iff $(k \supset \alpha)[j_1, \ldots, j_m, j]$ is.

We can now prove that RES$[\![k, \alpha]\!]$ captures within a first-order formula what it means for a KB whose knowledge is represented by k to know that α is true.

Lemma 3.6. *For any sentence k of $\mathcal{L}$, any formula α of $\mathcal{L}$ and any parameters $j_1, \ldots, j_m$, the sentence* $(\boldsymbol{K}\alpha \equiv \text{RES}[\![k, \alpha]\!])[j_1, \ldots, m_m]$ *is true on $\mathcal{R}[k]$.*

Proof. Suppose RES$[\![k, \alpha]\!][j_1, \ldots, j_m]$ is false on $\mathcal{R}[k]$. It cannot, therefore, be valid and so, by Lemma 3.5, neither can $(k \supset \alpha)[j_1, \ldots, j_m]$. So, there must be a world structure on which k is true and $\alpha[j_1, \ldots, j_m]$ is false. Thus, $\boldsymbol{K}\alpha[j_1, \ldots, j_m]$ is false on $\mathcal{R}[k]$. Conversely, assume that RES$[\![k, \alpha]\!][j_1, \ldots, j_m]$ is true on $\mathcal{R}[k]$. Since this sentence is pure and contains no $\boldsymbol{K}$-operators, it must be valid. So, again by Lemma 3.5, $(k \supset \alpha)[j_1, \ldots, j_m]$ must be valid also and, therefore, $\boldsymbol{K}\alpha[j_1, \ldots, j_m]$ must be true on $\mathcal{R}[k]$. To summarize, RES$[\![k, \alpha]\!][j_1, \ldots, j_m]$ is true on $\mathcal{R}[k]$ iff $\boldsymbol{K}\alpha[j_1, \ldots, j_m]$ is.

Lemma 3.6 completes the proof of the Representation Theorem.

3.2.3. *Remarks*

The proof of the Representation Theorem clearly depends on being able to finitely characterize the set of entities known to have some property whenever the knowledge of the KB is finitely represented. To appreciate the non-triviality of this property, it is worth examining a slight modification to $\mathcal{L}$ (called $\mathcal{L}'$) for which the property no longer holds and, consequently, for which the Representation Theorem is false.

There are, of course, first-order dialects other than $\mathcal{L}$ that *do* satisfy the Representation Theorem. In fact, a dialect without any parameters at all will do the trick but in a very trivial way. Without standard names, there is no way to pick out any specific individual. For example, telling a KB that $\phi(c)$ is true, where c is a 0-ary function symbol, is like telling it $\phi(f(g(c, d)))$: it knows that some individual is a ϕ, but not which. In fact, it can be shown that the only time a KB will have a *known* ϕ is when it knows that everything is a ϕ. So the sentence

$$\exists x \boldsymbol{K}\phi(x) \supset \boldsymbol{K}\forall x \phi(x)$$

will end up being a theorem. What this means, then, is that RES$[\![k, \alpha]\!]$ (where α has a single free variable) can be defined as $(x = x)$ or $(x \neq x)$, depending on k. Because the set of known instances of formulas is either everything or nothing, it can be characterized easily. This is enough to satisfy the Representation Theorem.[26]

Consider, on the other hand, a language $\mathcal{L}'$, like $\mathcal{L}$, except that it has a single parameter *1* and a special one-place *successor* function $'$ (written in postfix

notation), that can be applied to any term. Instead of having terms like

$$1, 2, 3, \ldots,$$

there are terms

$$1, 1', 1'', \ldots$$

which play the same role in the language. We will call these terms *pseudo-parameters*. If the successor function was only used for pseudo-parameters, $\mathscr{L}'$ would just be a notational variant of $\mathscr{L}$. What makes the difference is that the successor function can be applied to *any* term including variables.

We will leave the semantics and proof theory of $\mathscr{L}'$ undefined since most of it is inherited from $\mathscr{L}$. One difference is that the Axiom of Equality can now be stated directly (without using a schema) as

$$\forall x((x = x) \wedge (1 \neq x') \wedge \forall y[(x' = y') \equiv (x = y)]).$$

Sentences of the form

$$(t_1 = t_1) \wedge (t_1 \neq t_2),$$

where t_1 and t_2 are distinct pseudo-parameters, are all implied by the new axiom. Moreover, if Universal Generalization remains unchanged except for dealing with pseudo-parameters instead of parameters, the converse is also true.

The language $\mathscr{KL}'$ is the modification of $\mathscr{KL}$ that has the successor function. The important point is that this function has to be special in $\mathscr{KL}'$ the same way that equality and parameters are special. In particular, a sentence can still be pure even if it contains an occurrence of the successor function symbol. Similarly, the Axiom of Specialization is not restricted in the placement of a $'$ symbol within the scope of a **K**, as with other function symbols. What this amounts to is that just as a KB is assumed to know the truth value of $(t_1 = t_2)$ whenever it knows the referent of t_1 and t_2, it must know the referent of t' whenever it knows the referent of t. Another way of looking at this is to think

[26]As discussed in [7], this argument has more general implications for logics of knowledge and belief. It appears, for example, that if we assume that what an agent knows can be represented symbolically in a language (or, equivalently, that what he knows is the result of being told sentences in a language), then either there has to be rigid designators in that language or the kind of 'wh-knowledge' (or *de re* knowledge) he will have will be limited as above. In systems without rigid designators that do not have the above sentence as a theorem, there is a (perhaps implicit) assumption that wh-knowledge comes in some other form.

of parameters as *rigid designators* in that they always stand for the same things no matter how the other function and predicate symbols are interpreted (relative to an s and v). Similarly, equality is a *rigid predicate* in that it always stands for the identity relation. So the successor function is a *rigid function* since it too always stands for the same function. The key property about these symbols is that because they are effectively *logical* symbols, the KB must have complete knowledge of them. If it did not, pseudo-parameters could not be used to play the role of parameters.

However intuitively appealing this dialect may be, it cannot be used as a representation language since it is too expressive to satisfy the Representation Theorem. To see this, consider, to the contrary, the knowledge represented by

$$k = (\phi(1) \wedge \forall x[\phi(x) \supset \phi(x'')]) .$$

What are the known ϕ in this case? Certainly 1 is one. Moreover, since the KB knows that $\phi(1'')$ is true, $1''$ is a known ϕ also. In fact, every 'odd' pseudo-parameter is a known ϕ. The question then is how to represent in $\mathscr{L}'$ the result of

$$\text{TELL}[\![\mathscr{R}[k], \exists x[\psi(x) \wedge \mathbf{K}\phi(x))]]\!] .$$

What should RES$[\![k, \phi(x)]\!]$ be? There is a real problem here since instances of this formula must be either valid or unsatisfiable. Specifically, we cannot use something like number theory to characterize the odd pseudo-parameters since that would presumably involve using non-logical predicate or function symbols, thereby making the interpretation of the formula world dependent. It seems that the best we can do is use the infinite formula

$$(x = 1) \vee (x = 1'') \vee (x = 1'''') \vee \cdots .$$

The conclusion is that $\mathscr{L}'$ is too powerful a language since it allows a finitely represented KB to have a set of known ϕ that cannot be finitely characterized.

This does not mean that we cannot use $\mathscr{L}$ to provide a KB with a (finitely axiomatized) number theory of some sort. This theory may very well use a function symbol ′ to mean the successor function over numbers. Any such KB would obviously know all the implications of the theory.[27] In this situation, however, the successor function would not be rigid. The KB need not know how the numerals (as represented by repeated applications of the successor function to some initial parameter or constant) line up with the parameters (even if the theory stipulates that everything is a number). For instance, given the knowledge represented by k above (in addition to some number theory),

[27]It could not, however, know all the *truths* of number theory, as Gödel has shown.

the parameter *1* would be the *only* known ϕ even though the KB would know that there were infinitely many. Number theory would play a role no more special than a theory about cities and states.

The implication of the above argument is that even a small part of number theory (in particular, a theory of the successor function) cannot be incorporated into the *logic* of $\mathscr{L}$ without losing the Representation Theorem. In other words, knowledge resulting from being told sentences in $\mathscr{L}'$ augmented by the $\boldsymbol{K}$ operator *cannot* be represented in $\mathscr{L}'$ itself. This impossibility is not a result of computational intractability or even of undecidability; there is no oracle that could be defined that would help us represent the knowledge. It is purely a property of what $\mathscr{L}'$ can and cannot express. What all this suggests is that the language $\mathscr{L}$ is very close to the limit of expressibility that a first order dialect can have and still satisfy the Representation Theorem. The characterization of what changes are in fact possible remains an open problem.

4. Extensions and Applications

Having investigated some of the technical issues underlying a knowledge level view of a KB, we will now examine some possible extensions to the framework. Specifically, new knowledge level capabilities will be introduced in terms of extensions to the TELL and ASK operations. In the first subsection, we consider an alternative form of ASK that handles wh-questions. Next, we consider a new ASK (and TELL) that provides a knowledge level view of non-monotonic reasoning. Finally, we briefly consider a TELL that allows definitional mechanisms to be incorporated into the framework. Although all these preliminary extensions are in need of further research, they should at least suggest how the framework can be applied to some of the current research problems in knowledge representation.

4.1. Wh-questions and numeric questions

One of the original motivations behind the use of $\mathscr{KL}$ as a query language for a first-order KB was for those situations where wh-questions were problematical. We can think of a wh-question as a formula with free variables and the answer(s) as being the parameter substitutions that make the formula true. For example, the formula

MajorCity(city, state)

asks what the major cities and their states are. The problem with this question is that a KB may not have a list of major cities and states, but only general pieces of information about them. In this case, it is not even clear what the *form* of an answer should be.

A solution to this problem proposed in [13] is to consider wh-questions as

coming in two forms: the first form asks for *known* instances of the formula; the second form asks for *potential* instances of the formula.[28] Obviously, the known instances are a subset of the potential instances. But more importantly, as far as the KB is concerned, the *actual* instances fall somewhere between the known and the potential ones. This is a consequence of the fact that $\mathbf{K}(\mathbf{K}\alpha \supset \alpha)$ and $\mathbf{K}(\alpha \supset \neg\mathbf{K}\neg\alpha)$ are both theorems of $\mathcal{KL}$. So, as observed by Lipski, the known and potential instances of a formula place lower and upper bounds respectively on the set of actual instances. Moreover, unlike the actual instances of a formula, a KB (that has complete knowledge about itself) always knows what the known and potential instances are.

One thing to notice about this form of answers to wh-questions is that if formulas are taken from $\mathcal{KL}$, it is sufficient to deal with known instances. Instead of asking for known and potential instances of α, we instead simply ask for the known instances of $\mathbf{K}\alpha$ and $\neg\mathbf{K}\neg\alpha$ respectively. This suggests that a wh-question facility can be defined by

$$\text{Wh-Ask}[\![k, \alpha]\!] = \{\langle i_1, \ldots, i_n\rangle \mid \mathbf{K}\alpha[i_1, \ldots, i_n] \text{ is true on } k\}.$$

The question itself determines whether known or potential instances will be retrieved, with known instances as the default (for example, when formulas of $\mathcal{L}$ are used).

Before dealing with the issue of what to do when the set of substitutions is infinite (and therefore, cannot be 'returned' by Wh-Ask), it is worth considering the form of an answer when the set is finite. Typically, a user will not be interested so much in a set of parameters since these only represent equivalence classes and have no descriptive import. For example, if the answer to the question

MajorCity(city, California)

is the set $\{4, 7, 9\}$, not much is learned except that there are three cities. What is also required is way of moving from a vacuous designator to coreferential terms that have descriptive (or at least connotative) import. So the second facility that must be provided is a function

$$\text{DESCRIBE}[\![k, i]\!] = \{t \mid t \text{ is not a parameter and } \mathbf{K}(t = i) \text{ is true on } k\}.$$

Given a parameter, this function returns the set of all terms that are known to

[28]By a 'potential' instance, we mean a substitution of the free variables which is not known to be a negative instance of the formula.

be coreferential. The trouble, of course, is that there may be infinitely many such terms (even when the KB is representable).

There are a few ways of dealing with this. Because a KB can only mention finitely many function symbols, the terms have to use the same symbols repeatedly. The easiest way to define a finite DESCRIBE is to consider only primitive terms (that is, those with exactly one function symbol). For example, the parameter *4* above might be described by

$$\{\text{SanFrancisco}, \text{FavoriteCity}(3)\}\,.$$

Similarly, *3* might be described by

$$\{\text{FirstChild}(1, 6), \text{BestFriend}(5), \text{OldestBrother}(2)\}\,.$$

A simple brute force way of implementing this version of DESCRIBE would be to loop through all the function symbols used in the KB and, for each one, consider the application of the function to each tuple (of the arity of the function symbol) consisting of parameters mentioned in the KB (which are also finite in number), returning those that are known to be equal to the original parameter argument. Another way of dealing with the problem would be to return only those terms that use a function symbol fewer than some number of times (passed as an argument, perhaps). An ideal solution might be to devise a *grammar* for describing the set of terms to be returned. For example, the set

$$\{6, f(6), f(f(6)), \ldots\}$$

can be described by the grammar

$$\langle\text{solution}\rangle ::= 6 \mid f(\langle\text{solution}\rangle)\,.$$

It is not clear how powerful this kind of grammar would have to be.

Returning now to Wh-Ask itself, we are faced with the problem that the set of substitutions itself need not be finite (an obvious example being the answer to the question $x \neq y$). However, in this case, there is a very definite way to characterize the infinite set of parameter tuples. Specifically, for any KB represented by k, we can define

$$\text{Wh-Ask}[\![\mathscr{R}[k], \alpha]\!] = |\boldsymbol{K}\alpha|_k\,.$$

What this means is that the result (after simplifications) of Wh-Ask will be a formula like

$$(x = 2 \wedge y = 8) \vee (x = 4 \wedge y \neq 7)\,.$$

Determining actual instances given a formula of this kind is a completely straightforward process.

Before leaving the topic of extensions to ASK, it is worth considering a different variety of question: those dealing with the size of sets. In the same way a first-order KB is taken to have knowledge about what it does and does not know (without containing any explicit sentences to this effect), it makes sense to assume that a KB implicitly knows the size of certain sets. For example, if a KB knows that

$$\exists x \exists y((x \neq y) \wedge \forall z[\phi(z) \equiv (z = x \vee z = y)]),$$

then the set of known ϕ is empty and everything is in the set of potential ϕ so the two wh-questions are uninformative. However, the KB does know something very specific about the ϕ, namely, that there are exactly two of them. What would be convenient in this case is the ability to ask the KB how many instances of α there are (allowing *unknown* and *infinitely many* as answers). This does not seem to presume any special number theoretic abilities on the part of the KB other than the ability to count. In fact, we can *already* ask, for any given n, if there are (exactly, at most, at least) n instances by using existential quantifiers and inequalities. One can also imagine a slightly more general question which is to name the greatest known lower bound and the least known upper bound on the size of the set of instances. Even with no information at all, a KB could return $\langle 0, \infty \rangle$ as the interval and subsequently sharpen these bounds as knowledge is acquired. In fact, the knowledge the KB has about the size of the set can be arbitrarily vague. For example, a KB that knows the negation of the above sentence only knows that the number of ϕ there are is not two. So it might be worth developing a range of questions to find out exactly what a KB knows about the number of instances of some formula. On a different axis, when dealing with formulas with several free variables, we will want to consider questions that talk about mappings between the variables. For example, answers to questions like

How many y's per x satisfy $\alpha[x, y]$?

are much more informative than the size of the set of the tuples. Again, the range of questions to consider here is completely open ended.

4.2. Default reasoning

There is an interesting parallel between the way RES$[\![k, \alpha]\!]$ works and the circumscription mechanism of [14]. The parallel is that while circumscription is a formal method for conjecturing that the only entities satisfying a predicate are those that must satisfy it (according to some theory), RES$[\![k, \alpha]\!]$ is a method

of getting hold of those entities that are known to satisfy it. If we let $\vdash_\phi$ be the provability relation when the predicate ϕ is circumscribed, we get, for example,

$$(\phi(1) \wedge \phi(2)) \vdash_\phi \forall x(\phi(x) \equiv (x = 1 \vee x = 2)).$$

In other words, if *1* and *2* are ϕ, then (if the only ϕ are those that must exist), *1* and *2* are the only ϕ. In our case, we can start by defining a provability relation $:\vdash$ by

$$k :\vdash \alpha \quad \text{iff} \quad \vdash (k \supset |\alpha|_k)$$
$$\text{where } k \in \mathscr{L} \text{ and } \alpha \in \mathscr{KL}.$$

One reading of this non-monotonic operator is

> If k represents everything that is known,
> then α is known to be true.

So here, for example, we have that

$$(\phi(1) \wedge \phi(2)) :\vdash \forall x(\boldsymbol{K}\phi(x) \equiv (x = 1 \wedge x = 2)).$$

This says that if it is known that *1* and *2* are ϕ, then if nothing else is known, *1* and *2* are the only *known* ϕ. Note that unlike circumscription, there is no conjecture here; there may very well be other instances of ϕ, but *1* and *2* are indeed the only known ones. However, we could get exactly the same result as circumscription by conjecturing that every ϕ is known, that is by adding the *closed world assumption* [15] for ϕ.[29] In this case, the ϕ would be identified with the known ϕ and the KB would then believe that *1* and *2* are the only ϕ. In other words, $\mathscr{KL}$ allows us to divide circumscription into two components: first, identify the known instances of a predicate, then conjecture that every instance is known.

This is not to say that there is an *exact* correspondence between circumscription and RES$[\![k, \alpha]\!]$ (with ϕ replaced by $\boldsymbol{K}\phi$). It appears that neither formal system can simulate the other. For example, with a theory containing only $\exists x\phi(x)$, circumscription would conclude that there was a *single* ϕ, while there are no known ϕ at all. On the other hand, the similarity suggests that $\mathscr{KL}$ might be a reasonable place to consider formulating yet another account of non-monotonic reasoning [16]. In particular, reasoning by default [17] seems to consist of making assumptions *in the absence of knowledge to the contrary.* It is this auto-epistemic aspect that is not addressed in other formalizations and that makes the language $\mathscr{KL}$ especially relevant.

[29]An important property of $\mathscr{KL}$ is that this closed world assumption can be expressed as a sentence: what we want to TELL the KB is that $\forall x(\phi(x) \supset \boldsymbol{K}\phi(x))$.

The idea behind the treatment of defaults suggested in [7] is the following. Consider two query operators: ASK, as before, and ASK* which answers taking defaults into account. For example, given a KB whose knowledge is represented by Bird(*1*), we would have that

$$\text{ASK}[\![\mathscr{R}[\text{Bird}(1)], \text{Fly}(1)]\!] = \text{unknown} ,$$

but

$$\text{ASK}*[\![\mathscr{R}[\text{Bird}(1)], \text{Fly}(1)]\!] = \text{yes} .$$

In other words, ASK* is willing to assume by default that the bird in question flies.

This effect can be defined precisely in terms of an ASK* that behaves exactly like ASK except with respect to a KB that has some additional beliefs about birds. In other words, for some δ, ASK* here could be characterized[30] by

$$\text{ASK}*[\![k, \alpha]\!] = \text{ASK}[\![\text{TELL}[\![k, \delta]\!], \alpha]\!] .$$

The δ should state that 'by default' birds fly. One way to paraphrase this, given $\mathscr{KL}$, is to state that all birds fly except those known not to fly:

$$\delta = \forall x([\text{Bird}(x) \wedge \neg \boldsymbol{K} \neg \text{Fly}(x)] \supset \text{Fly}(x)) .$$

An alternative formulation might be to say that all *known* birds fly except those known not to fly (that is, the first predicate in δ would be $\boldsymbol{K}\text{Bird}(x)$). One difference between the two is that in the original formulation only, with no information at all, a KB would assume that all birds fly. Other differences are discussed in [7].

A sentence of this type might be called a *default*: a statement that certain individuals have a property provided that they are not known not to have the property. So like the formalism of [18] but unlike that of [19], defaults are sentences of a language, that is, potential objects of belief which have a very natural proof theoretic *and* semantic characterization by virtue of the Representation Theorem. Moreover, unlike both of these approaches, ours does not require defining a logic with new inference rules[31] or sentential operators. In fact, the sentential operator $\boldsymbol{M}$ of [18] seems to be related to $\boldsymbol{K}$. Their interpretation of $\boldsymbol{M}\alpha$ is that α is *consistent* with what is believed. However, it is not clear how important it is to interpret the symbol in terms of what can or cannot be *derived* in a formal system. It seems more appropriate to interpret it (epistemically) as saying that α is not *known* to be false, that is, as $\neg \boldsymbol{K} \neg \alpha$.

There is another approach to defaults we might consider, which is to leave ASK unchanged but provide a new function TELL* that behaves like TELL

[30]It is important to realize that this definition of ASK* only states what the behavior of defaults must be and not how they could be *implemented* at the symbol level.

[31]See [12] for arguments about why non-monotonic belief revision should not be construed as following a logic with non-monotonic inference rules.

but taking defaults into account. The difference between the two methods is that ASK* answers questions by first making some assumptions while leaving the KB unchanged; TELL*, on the other hand, would actually change the KB allowing the default to become a regular belief. In other words, ASK* says that defaults are assumptions that can be made to answer a certain form of question, while TELL* says that defaults are beliefs that are acquired by virtue of what is left unsaid in a certain form of assertion. This distinction is difficult if not impossible to capture within existing formalizations of default reasoning but is handled quite naturally in our functional framework.

Obviously, a general definition of ASK* (or TELL*) should not depend on the default properties of birds specifically. A more reasonable definition that depends only on the query and the current KB is presented in [7]. Essentially, the idea is to separate a default into two pieces. The first involves knowledge about typical birds, for example, that they fly, have two wings, and so on. A new predicate forming operator ∇ (read as '*typical*') is introduced so that if "Bird" is a one place predicate, then so is '∇Bird'.[32] So, for example, a KB may know that

$$\forall x(\nabla\text{Bird}(x) \supset \text{Fly}(x)) .$$

without having any idea of which birds are typical (although it will realize that any bird that does not fly is atypical). This is not a belief about all birds in general, just the typical ones. But the crucial points is that this is *not* default knowledge: the belief here is that typical birds fly—not that typical birds *typically* fly. In fact, we may want to say something stronger than this, namely, that flying is not only not a default property of typical birds, it is an *essential* property. The impact of this distinction will be taken up again below.

The second part of a default says which birds are typical and is the sentence that will be assumed by ASK*:

$$\forall x([\text{Bird}(x) \wedge \text{—}\boldsymbol{K}\text{—}\neg\nabla\text{Bird}(x)] \supset \nabla\text{Bird}(x)) .$$

In other words, the assumption is that any bird not known to be atypical is typical. Although this sentence uses the predicate 'Bird', it would have exactly the same form for any other predicate. In fact, δ can be defined (almost) as the conjunction of sentences of this form over all predicates used in the KB (except those preceded by a ∇).[33] Returning to our above example, the end result is

[32]Obviously we want ∇ to apply to n place predicates for any n. Moreover, because a bird can be typical with respect to number of legs but atypical with respect to flying ability, there will be a family of predicates formed by ∇, given a predicate and an index.

[33]The simple conjunctive definition has to be modified because of the complications that arise when trying to apply multiple competing defaults simultaneously. These issues are discussed in [20] and [7].

that while the original KB will not know if the bird flies, it will believe that if it is typical, then it flies. Moreover, the augmented KB that is used for ASK* will also assume that it is typical (since there is no knowledge of it being atypical) and, therefore, that it flies.

To summarize, the notion of default that seems the best suited to $\mathcal{KL}$ is a very simple one where certain instances of predicates are assumed to be typical. All of the 'content' of the default is put into knowledge about the properties of typical instances of the predicates. While much remains to be done on the semantics of the ∇-operator, there is already evidence from other sources that it is a useful abstraction [21].

4.3. Definitional mechanisms

In terms of system design, the main reason for distinguishing between the knowledge level and the symbol level is to allow the functionality of a system to be treated independently of its symbolic implementation. In particular, it allows us to consider new operations on a KB (that can be explained in terms of existing ones) without necessarily committing ourselves to any particular implementation style.

Consider an operation that allows new non-logical terms to be *defined* in terms of existing ones.[34] For example, we may decide to start using a predicate 'CapitalCity' and want it to apply only to cities that are state capitals. One way of informing a KB of this is to assert (using TELL) something like

$$\forall x(\text{CapitalCity}(x) \equiv \text{City}(x) \wedge \exists y(\text{State}(y) \wedge \text{Capital}(y) = x)) .$$

However, even though this assertion gets the KB to believe the right things about capitals, it does not tell it that this is how the term is being defined. To see this, suppose we just wanted to tell the KB that (it so happened that) someone called Joe liked every city except the capitals. We could then justifiably assert

$$\forall x(\text{CapitalCity}(x) \equiv \text{City}(x) \wedge \neg\text{Like}(\text{Joe}, x)) .$$

Notice that the form of these two sentences is very similar; if either one is supposed to be a *definition* of 'CapitalCity', then they should both be definitions. But we certainly do not want to *define* capital cities as those that Joe does not like. Nor, for that matter, do we want to say that every capital city just so happens to be a capital of some state, as if it were completely by accident or even possible that some were not. Indeed, the trouble with trying to identify definitions in terms of the logical form of certain assertions is that not only are some universally quantified biconditionals not definitional, there may

[34] See [4] for arguments against the prevalent position that definitions are only relevant in formal domains (such as mathematics).

very well be definitions of terms that are not universally quantified biconditionals.

It should be obvious that our TELL operation does not capture the difference in intent between the two sentences. Its purpose is to allow a KB to be told true sentences without regard to whether the truth is based on a property of the world or on the definition of a term. As it stands, there is no way to *define* a term like 'CapitalCity' so as to cause sentences like the first to be believed. To the extent that this kind of functionality is deemed useful in the design of knowledge-based systems (see [22], for example), a new operation is required.

The idea here is to have an operation

DEFINE[[k, term, definition]]

that takes a KB, a predicate (or function) symbol and a definition for this predicate and has as value a KB that knows how the term is defined. The actual form of a definition is not really relevant here, though what we have in mind is a small number of *predicate forming operators* that allow complex predicates to be formed out of simpler ones. For example, a predicate forming operator Qua could be used to define capitals as those cities that play the role of state capital for some state:

DEFINE[[k, CapitalCity, (Qua City Capital State)]].

The only major question to answer here is what the definition of the DEFINE operation should be. Perhaps the simplest solution is to treat a definition exactly as before, that is, as synonymous with an assertion of some sentence, a *meaning postulate* associated with the term forming operator. For example, we could understand an operation like

DEFINE[[k, ϕ, (Qua $\psi\,\xi\,\theta$)]]

as if it were

$$\text{TELL}[[k, \forall x(\phi(x) \equiv \psi(x) \wedge \exists y[\theta(y) \wedge \xi(y) = x])]].$$

This form of 'macro'-definition performs exactly the same assertion as an explicit TELL except that we have separated the actual term formation from the meaning postulate.

Even though we can explain a DEFINE in terms of a TELL, there are good reasons for keeping the two separate. First of all, the two operations are different in intent and thus have different standards of adequacy. For example, the adequacy of the TELL operation is measured in terms of its ability to form very weak assertions. Incompleteness is not an issue with term formation, however. As argued in [22], what is likely to count in a definitional facility is

the ability to generalize and specialize existing terms and to aggregate collections of terms into larger complex wholes.

Moreover, the distinction between the two operations encourages us to consider implementation methods for definitions that are not based (in the obvious way) on the implementation used for the TELL operation. For example, under the (very plausible) assumption that it will be convenient to move quickly from a term to its definition, we might consider using an inheritance network data structure to keep track of relationships among the various terms being defined. In this sense, a definition allows us to inform the symbol level about what sentences of a certain form are important enough to be treated specially, something that would not be possible if only assertions were allowed.

Finally, even without considering any additional operations, we can present an interface to a KB that appears to extend beyond the first-order capabilities. For example, we might allow

$$\text{DEFINE}[\![k, \phi, (\text{TransitiveClosure } \psi)]\!]$$

to mean

$$\text{TELL}[\![k, \forall x \forall y (\phi(x, y) \equiv (x = y) \vee \exists z [\psi(x, z) \wedge \phi(z, y)])]\!] .$$

This would, for example, allow us to define 'ancestor' directly as the transitive closure of 'parent' without having to characterize a transitive closure relation over predicates using a second order theory. As it turns out, transitive closures also happen to be predicates that are best implemented at the symbol level in terms of special-purpose inference techniques.

If we consider an extension to our framework that allows new *query* operations, the difference between defining a term and asserting a sentence becomes more substantial. For example, if we can ask if one term analytically subsumes[35] another, then we can obviously notice the difference between defining a term in a certain way and asserting some corresponding sentence (meaning postulate). The same is true if we can access the definition of a term directly. This kind of capability allows us to examine how a predicate was defined independently of what is known about (instances of the) predicate. In a sense, our functional view of a KB is enlarged to reflect a distinction between *essential* properties of instances of a term (arising by virtue of the way the term was defined) and *factual* ones (arising by virtue of what is true in the world).

The issue here, of course, is how to define these extended operations. It is no longer sufficient, for example, to explain DEFINE in terms of a TELL if we have to be able to calculate subsumption relationships. The simplest solution

[35] A predicate analytically subsumes another if, by virtue of how the predicates are defined, every instance of the latter predicate must also be an instance of the former.

might be to divide a KB into two parts, the first as before, and the second containing the effect of the definitions. The regular ASK operation would use both parts of the KB while analytic questions would only consider the second part. This second component could be a set of world structures, or (syntactic) definitions of some sort, or even meaning postulates. A more radical possibility is to make the definitional component incorporate special relations over predicates that are part of the semantics of the language itself. This is essentially what is done in [23], where a lattice structure over predicates (where the partial ordering is that of conceptual containment) is part of the model theory of a first-order language.

Another possible extension to our framework is to allow for the definition of predicates which cannot be characterized in terms of necessary or sufficient conditions. For such terms (like those standing for *natural kinds*) any 'definition' is best thought of in terms of *prototypical* rather than *essential* properties. These are properties that instances of the term might be *expected* to have. If we take these prototypical properties to be part of how the predicate is defined, a simple approach suggests itself. Basically, we can use the default mechanism discussed above and consider the definition of any predicate ϕ to include the definition of $\nabla\phi$. So, to make *flying* be a prototypical property of birds, we make it be an essential property of *typical* birds. In other words, the prototypical properties of 'Bird' are just the essential properties of '∇Bird'. Overall, then, instances of a predicate would be characterized in terms of factual, essential, and prototypical properties (leaving open, for the moment, the epistemological status of the factual properties of instances of '∇Bird').

This simple expedient allows us to consider definitions of terms that include both essential and prototypical features. In other words, we can support the kind of 'definition' found in frame-based representation languages like UNITS [24] or KRL [25] (modulo syntactic sugar, of course). But most importantly, because we have separated the knowledge level from the symbol level, we can *implement* these definitions in any convenient way that preserves the semantics of the operations. Specifically, we are free to use the same methods found useful in the frame-based languages (such as inheritance networks), rather than more general theorem proving techniques better suited to TELL and ASK.

To summarize, if we extend the framework to incorporate definitional facilities and especially if we include default capabilities, we can provide a framework that has aspects of both logical and object-oriented representation systems. As such, it offers an especially convenient formal starting point for the development of 'hybrid' systems like the kind discussed in [22, 23, 26].

5. Conclusion

In this paper, we have investigated some of the technical properties of a functional approach to knowledge representation where *what* a KB does for a

system is kept quite separate from *how* it represents what it knows. This, we argued, was in keeping with current programming principles (exemplified in the abstract data type methodology), as well as being compatible with Newell's notion of knowledge and symbol levels.

Our first concern was to develop a language for *interacting* with a KB, a first order logical language called $\mathscr{L}$. Because the usual role of logical languages in AI has been to store information (at the symbol level), our dialect had to have special features such as parameters and a non-standard treatment of equality. We illustrated the appropriateness of this dialect by showing that a slightly stronger one (called $\mathscr{L}'$) was too expressive to satisfy the Representation Theorem.

We then showed that using $\mathscr{L}$ as an assertion language can result in an incomplete KB that is best queried in a more powerful language. This led to a generalization of $\mathscr{L}$, called $\mathscr{KL}$, which allows reference to not only the domain, but to the KB's knowledge of the domain, its knowledge of its knowledge of the domain, and so on. We gave a formal semantics for $\mathscr{KL}$ independent of the one for $\mathscr{L}$, something we could not have done if $\mathscr{KL}$ had merely been a first-order theory (using syntactic encodings of sentences).

The important property of $\mathscr{KL}$ as an interaction language is that it allowed closed (or open) world assumptions and defaults, among other things, to be expressed as sentences of the language. Since TELL and ASK were defined in terms of $\mathscr{KL}$, a KB could be informed or queried about these sentences. The result is that, unlike other approaches, a KB could be made to believe closed world assumptions selectively like any other knowledge about the domain.

The major technical result of this research, the Representation Theorem, establishes that even with the extended expressive power provided by $\mathscr{KL}$, the knowledge of a KB is still representable at the symbol level using sentences of $\mathscr{L}$. In other words, the theorem shows the *correctness* of a first order symbolic realization of a KB. Again, had $\mathscr{KL}$ been specified as a first-order theory, no such proof would have been possible since the behavior of a KB would have been *defined* in terms of its first-order implementation.

In the final section, we argued for the robustness (and relevance) of our approach by showing how it could be applied to some current knowledge representation problems. The danger in a formal logical investigation like the one we have undertaken is that results may be regarded as 'merely' formal or mathematical, with no relevance to real knowledge-based systems. Our feeling is that the functional viewpoint does have *practical* advantages which we are currently attempting to exploit in a knowledge representation system called KRYPTON [27]. The two main features of KRYPTON are an interface based on TELL and ASK and a division of the functionality into definitional and assertional components along the lines discussed above.

Once knowledge level concerns have been separated from those at the symbol level, the kind of issues discussed here become very relevant to the

design of knowledge representation systems. The functional approach we have considered applies equally well to languages that are less expressive than $\mathcal{KL}$. Without advocating a specific position on the expressiveness *vs.* computational tractability trade-off, the framework offers a formal foundation for comparing and evaluating competing proposals by concentrating on the service provided by a KB independently of how this service is realized. Once the *competence* of the knowledge representation system has been established, its *performance* characteristics can be examined in proper perspective.

ACKNOWLEDGMENT

This research is based on my doctoral dissertation at the University of Toronto. I want to thank my thesis supervisor, Prof. John Mylopoulos, for his invaluable moral and technical support. I also wish to acknowledge my colleagues at the University of Toronto, BBN, and FLAIR for their many contributions to the ideas reported here and to thank in particular Ron Brachman, Phil Cohen, Joe Halpern, Gerhard Lakemeyer, Peter Patel-Schneider, Jim Schmolze, and an anonymous referee for their comments on an earlier draft of this manuscript. Financial support was gratefully received from the Department of Computer Science of the University of Toronto and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

1. Newell, A., The knowledge level, *AI Magazine* **2**(2) (1981) 1–20.
2. Liskov, B. and Zilles, S., Programming with abstract data types, *SIGPLAN Notices* **9**(4) (1974).
3. Gallaire, H. and Minker, J. (Eds.), *Logic and Data Bases* (Plenum, New York, 1978).
4. Israel, D.J. and Brachman, R.J., Distinctions and confusions: A catalogue raisonne, *Proc. IJCAI-81*, Vancouver (1981) 452–459.
5. Mendelson, E., *Introduction to Mathematical Logic* (Van Nostrand–Reinhold, New York, 1964).
6. Levesque, H.J., The interaction with incomplete knowledge bases: A formal treatment, *Proc. IJCAI-81*, Vancouver (1981) 240–245.
7. Levesque, H.J., A formal treatment of incomplete knowledge bases, Tech. Rept. No. 3, Fairchild Laboratory for Artificial Intelligence Research, Palo Alto, CA, 1982.
8. Leblanc, H., On dispensing with things and worlds, in M.K. Munitz (ed.), *Logic and Ontology* (New York University Press, New York, 1973) 241–259.
9. Levesque, H.J., A formal treatment of incomplete knowledge bases, M. Brodie, J. Mylopoulos and J. Schmidt (Eds.), in: *Perspectives in Conceptual Modelling* (Springer, Berlin, 1984).
10. Hintikka, J., *Knowledge and Belief: An Introduction to the Logic of the Two Notions* (Cornell University Press, Ithaka, NY, 1962).
11. Moore, R., Reasoning about knowledge and action, Tech. Note 181, SRI International, Menlo Park, CA, 1980.
12. Israel, D.J., What's wrong with non-monotonic logic, *Proc. AAAI-80*, Stanford, CA, (1980) 99–101.
13. Lipski, W., On semantic issues connected with incomplete information data bases, *TODS* **4**(3) (1978).
14. McCarthy, J., Circumscription—A form of non-monotonic reasoning, *Artificial Intelligence* **13** (1980) 27–39.
15. Reiter, R., On closed world data bases, in: H. Gallaire and J. Minker (Eds.), *Logic and Data Bases* (Plenum, New York, 1978), 55–76.
16. Bobrow, D.G. (Ed.), *Artificial Intelligence* **13**, Special Issue on Non-Monotonic Reasoning, 1980.

17. Reiter, R., On reasoning by default, *Proc. Second TINLAP Conference*, Urbana, IL (1978) 210–218.
18. McDermott, D. and Doyle, J., Non-monotonic logic I, *Artificial Intelligence* **13** (1980) 41–72.
19. Reiter, R., A logic for default reasoning, *Artificial Intelligence* **13** (1980) 81–132.
20. Reiter, R. and Criscuolo, G., Some representational issues in default reasoning, Tech. Rept. 80–7, Dept. of Computer Science, University of British Columbia, Vancouver, 1980.
21. Cohen, B., Understanding natural kinds, Ph.D. Thesis, Dept. of Philosophy, Stanford University, Stanford, CA, 1982.
22. Brachman, R.J. and Levesque, H.J., Competence in knowledge representation, Proc. *AAAI-82*, Pittsburgh, PA (1982) 189–192.
23. Israel, D.J., On interpreting network formalisms, *Internat. J. Comput. Math.* **9**(1) (1983) 1–14.
24. Stefik, M., An examination of a frame-structured representation system, *Proc. IJCAI-79*, Tokyo (1979) 845–852.
25. Bobrow, D.G. and Winograd, T., An overview of KRL: A knowledge representation language, *Cognitive Sci.* **1**(1) (1977) 3–46.
26. Rich, C., Knowledge representation languages and predicate calculus: How to have your cake and eat it too, *Proc. AAAI-82*, Pittsburgh, PA (1982) 193–196.
27. Brachman, R.J., Fikes, R.E. and Levesque, H.J., KRYPTON: A functional approach to knowledge representation, *IEEE Comput.* **16**(10) (1983) 67–73.

Received March 1983

A Computational Theory of Belief Introspection

Kurt Konolige
Artificial Intelligence Center
SRI International
Menlo Park, California 94025

Abstract

Introspection is a general term covering the ability of an agent to reflect upon the workings of his own cognitive functions. In this paper we will be concerned with developing an explanatory theory of a particular type of introspection: a robot agent's knowledge of his own beliefs. The development is both descriptive, in the sense of being able to capture introspective behavior as it exist; and prescriptive, in yielding an effective means of adding introspective reasoning abilities to robot agents.

1 Introduction

Introspection is a general term covering the ability of an agent to reflect upon the workings of his own cognitive functions. In this paper we will be concerned with developing a theory of a particular type of introspection: an agent's knowledge of his own beliefs. There are at least two reasons why it is important to develop such a theory, one descriptive and the other prescriptive. As Collins and his coworkers have shown (in [1]), an agent often reasons about his own beliefs and nonbeliefs in deciding the answer to a posed query; hence a descriptively adequate account of agents' beliefs must deal with introspection. The second reason is that researchers attempting to build artificial

This research was made possible in part by a gift from the System Development Foundation. It was also supported by Grant N00014-80-C-0296 from the Office of Naval Research.

agents must imbue these agents with introspective knowledge if they are to act in an intelligent manner. Moore [11] gives the example of an agent who must introspect about his beliefs in order to form a correct plan to achieve a goal.

In this paper we offer an explanatory theory of belief introspection based on the concept of a *belief subsystem* as developed in Konolige [4], [5]. Put simply, a belief subsytem is the computational structure within an artificial agent responsible for representing his beliefs about the world. Because the belief subsystem is "at hand" and available to the agent, it is possible for the agent to gain knowledge of his beliefs by simply making recursive calls to this belief subsystem, perhaps with ever-decreasing resource allocations. This, in a nutshell, is the model of introspection we adopt. Its advantages are that it is an adequate explanatory theory of belief introspection, and that it is immediately prescriptive: the theory shows how artificial agents that exhibit introspective reasoning of the requisite sort can be built.

Given the importance of introspective reasoning, it is perhaps surprising that the problem of finding a good explanatory basis for belief introspection in artificial agents has scarcely been addressed. In Section 3 we review two approaches that differ from ours in being nonconstructive: an ideal agent's introspective reasoning is defined by putting constraints on her belief set. The disadvantage of such nonconstructive theories is that, in general, they do not extend to the case where an agent's reasoning powers are bounded by resource limitations.

2 The Introspective Machine

We start developing a theory of belief introspection by considering the computational embodiment of belief in an artificial agent. We have argued elsewhere (*e.g.*, Konolige [4]) for the identification of a *belief subsystem* as a conceptually separate part of an agent's cognitive makeup. A belief subsystem **M** consists of a finite list of facts the agent believes to be true of the world (the *base set*), together with some computational apparatus for inferring

consequences of these facts. **M** interacts with other cognitive systems of the agent (*e.g.*, a planning system) as a query-answering device. It accepts a query ϕ and attempts to show that ϕ be derived from its base set of facts. The ***belief set*** of an agent is the set of all queries that can be derived.

The queries presented to **M** are in an internal language L; the exact nature of this language is not important, but there must be expressions in it that refer to the agent's own beliefs. We take these expressions to be of the form $\Box\phi$, meaning the agent believes ϕ to be one of his own beliefs. Formulas of L not containing $\Box$ are called *nondoxastic*; the sublanguage of L consisting of all nondoxastic sentences is called the ***underlying language***.

When presented with a query in the language L, we assume **M** operates by either matching the query against its base set, or applying inference rules to generate subqueries in a backward-chaining manner. For example, in trying to answer the query $P \vee Q$, it may split the disjunc-

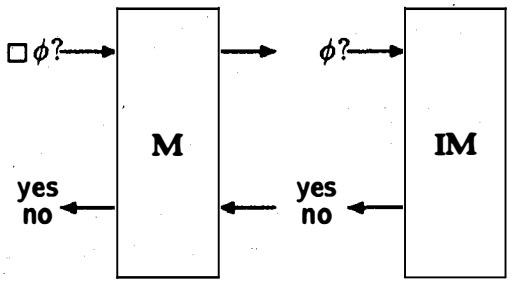


Figure 1: An Introspective Belief Subsystem

tion into two separate queries, and try to answer each of these. During the course of generating subqueries, it come upon one that is a question about its own state, *i.e.*, of the form $\Box\phi$. Such a subquery can be answered by making a recursive call to the belief subsystem again, posing the query ϕ. Conceptually we can think of this recursive call as a call on a new belief subsystem **IM** (the ***introspective machine***). The **IM** may have different characteristics than

M — for instance, it may have only a subset of the facts available to **M**, or even have facts that contradict those in **M** (according to Hintikka [3], people can have introspective beliefs of this sort). If the **IM** must answer a query about *its* self-beliefs, then it relies on another machine, the **IIM**; and so on, creating a hierarchy of belief subsystems. We write $\mathbf{I}^n\mathbf{M}$ to indicate the nth element of the hierarchy, with $\mathbf{M} = \mathbf{I}^0\mathbf{M}$. A belief subsystem that relies on an introspective machine to answer queries about self-beliefs in this manner is called an ***introspective belief subsystem.***

A query ϕ that is answered affirmatively by **IM** means that the agent, upon introspecting on his own beliefs, comes to the conclusion that she believes ϕ — that is, $\Box\phi$ is one of her beliefs. A negative answer, on the other hand, means that she doesn't believe her belief subsystem computes ϕ, and so in this case $\neg\Box\phi$ is one of her beliefs. Figure 1 illustrates the workings of the introspective machine by showing the way in which **M** responds to the query $\Box\phi$. **M** poses the subquery ϕ to the introspective machine. If **IM** answers *yes*, then $\Box\phi$ is accepted as a belief, and **M** also answers affirmatively. If **IM** answers **no**, $\Box\phi$ is not a belief.

There are no restrictions on the inference rules that a belief subsystem uses, except that they should be *sound* with respect to the semantics of the underlying language (they need not be complete). In particular, we wish to exclude rules of an introspective nature, because we want all properties of introspection to arise from the interaction of **M** and **IM**. For example, if the underlying language is propositional, the rules should respect the truth-functional semantics of the boolean connectives. In the case of a first-order underlying language, rules such as those in Kripke [6] or Konolige [4] may be used, disallowing the rules explicitly dealing with modal operators.

Proposition 1 *Suppose* **M** *is an introspective belief subsystem whose base sentences are nondoxastic and consistent. Then* **M** *is consistent.*

The proof[1] of this proposition follows from noting that

[1] Space requirements preclude more than sketches of most proofs in this paper.

M is atomically consistent. For nondoxastic atoms P, at most one of P or $\neg P$ will be derivable from the base sentences. For atoms $\Box\phi$, the responses given in Figure 1 indicate that either $\Box\phi$ or $\neg\Box\phi$ will be provable, but not both. Note that if **IM** is inconsistent, $\Box\phi$ and $\Box\neg\phi$ will be provable in **M** for some sentence ϕ.

2.1 Ideal Agents

An ideal agent should have perfect knowledge of her own beliefs. This motivates the following definition.

Definition 1 *An* ideal introspective belief subsystem **M** *satisfies three criteria:*

1. *The belief set of* **M** *is consistent.*
2. *The inference rules are complete.*
3. $\mathbf{I}^n\mathbf{M} = \mathbf{M}$ *for all* $n > 0$.

The first condition is that an ideal agent's beliefs not be contradictory. An interesting case of inconsistency occurs when the base set of **M** contains doxastic sentences. For example, suppose the base set of consists solely of the sentence $\Box P$. Now the query $\Box P$ will be answered affirmatively in **M** (by direct matching). The query $\neg\Box P$ can also be proven, because it generates the query P to **IM**. Since **IM** has the same base set as **M**, it answers P negatively, and so **M** affirms $\neg\Box P$.

Not all doxastic base sentences lead to inconsistency, of course; sometimes their presence is required for useful inference. Moore [12] gives the following example of an agent's introspective reasoning: "I don't have any brothers, because if I did, I would know about them, and I have no such knowledge." If we let P stand for "I have no brothers," then the agent's base set includes the axiom $\neg\Box P \supset \neg P$. Now given the query $\neg P$, the agent's belief subsystem would use the axiom and set up the goal of proving $\neg\Box P$. This generates a query P for the **IM**, which answers negatively. Hence $\neg\Box P$ is proven, and so is the original query $\neg P$.

The completeness of the inference rules is with respect to the semantics of the underlying language. If the rules are complete, the belief set is closed under the appropriate notion of logical consequence for the underlying language. In the case of a propositional language, it is closed under truth-functional consequence.

The third condition is the requirement that ideal agents have perfect introspective knowledge. We enforce this by assuming that an agent's view of her own belief subsystem (**IM**) is exactly the same as the real subsystem (**M**). Because the introspective machine has its own introspective machine (**IIM**), this too must be the same as (**M**), and so on to arbitrary introspective levels.

Proposition 2 *In an ideal introspective belief subsystem,*

PI: *If* **M** *responds* yes *to* ϕ, *it responds* yes *to* $\Box\phi$.

NI: *If* **M** *responds* no *to* ϕ, *it responds* yes *to* $\neg\Box\phi$.

To prove this proposition, note that the **IM** has exactly the same structure as **M**, and so has exactly the same behavior on a query ϕ. *PI* abbreviates *positive introspection*; informally, it says that if an agent believes a proposition, she believes that she believes it. Similarly, *NI*, or *negative introspection*, says that an agent has knowledge of what she does not believe.

In the best of all possible worlds, we could actually implement such an introspective belief subsystem, and so provide artificial agents with an ideal mechanism for reasoning about their own beliefs. Unfortunately, except under fairly strict conditions on the underlying language, there does not exist any realizable computational structure that will implement an ideal introspective belief subsystem.

Definition 2 *A belief subsystem* **M** *is* decidable *if there exists an algorithm that will return* yes *when* ϕ *is derivable and* no *when not; it is* semi-decidable *if there exists an algorithm which will return* yes *when* ϕ *is derivable; it is* undecidable *if it is not semidecidable.*

When we talk about the decidability of an introspective belief subsystem **M**, we normally take derivability to include the derivation of introspective beliefs via **IM**. Sometimes, however, we want to refer to the decidability of **M** without the introspective rules; to make this clear, we say "the decidability of nonintrospective **M**."

Proposition 3 *Let* **M** *be an ideal introspective belief subsystem. If nonintrospective* **M** *is not decidable,* **M** *is undecidable.*

The proof is simple: suppose **M** is semidecidable. Then there is an algorithm for determining that **M** returns **yes** on $\neg\Box\phi$ and $\Box\phi$, where ϕ is an arbitrary sentence. Thus there is an algorithm for determining whether ϕ is derived or not by **IM**, contradicting the assumption that **IM** is not decidable.

Proposition 4 *If the underlying language is propositional, and its base set is nondoxastic, an ideal introspective belief subsystem is decidable.*

The proof here is straightforward: any query will have a finite maximum embedding n of self-belief operators. One need only look at the (decidable) theorems produced by the first n levels of the introspective machine. As long as queries do not include any quantification into the context of the self-belief operator, we can extend this result to any underlying language which can be decided by reduction to the propositional calculus. For example, monadic predicate calculus (PC) and the class of $\exists\forall$-sentences have this property.

These two propositions to some extent delimit the nature of decidability for introspective subsystems. A natural question to ask is if Proposition 4 can be extended to the case of *any* decidable underlying language. The answer to this has important consequences for adding introspective ability to artificial agents, because these agents are (nonintrospectively) decidable: they must answer a belief query in a finite amount of time.

Proposition 5 *If the underlying language is monadic PC, and its base set is nondoxastic, an ideal introspective belief subsystem is decidable.*

The proof of this proposition relies on Kripke's result in [7] that monadic modal PC is not decidable. The difference between monadic modal PC and propositional modal languages is that the former allows quantifying into the modal context. As we mentioned, queries without quantifying-in are decidable for monadic PC. Thus the presence of quantifying-in seems to pose an inherently difficult computational problem for introspective systems. Yet the expressivity of quantifying-in is desirable in many applications; Levesque [9] gives the example of a question-answering system in which sentences of the form $\exists x[\, P(x) \wedge \neg \Box P(x)\,]$ express the fact that there are individuals with property P whose identity is unknown to the database.

Proposition 5 is discouraging, since it means that in constructing introspective agents, we must either use a very weak underlying language, or give up some of the three conditions of ideality. We discuss the latter method in the next section. Note that even without Proposition 5, there are reasons for developing the theory of non-ideal agents. First, even with a very weak underlying language and a decidable subsystem, an agent may have limited resources for derivation of beliefs, and can only compute an approximation to the conditions of Definition 1. Second, we mentioned that human agents are not always ideal agents, and we would like to model their cognitive behavior.

2.2 Real Agents

In Figure 1, a belief subsystem had to respond either **yes** or **no** to every query. In a computational setting with finite resource bounds, it may not be possible to do this in a consistent way. For example, if the underlying language is PC, there are some (nondoxastic) queries that do not have a derivation, and hence the belief subsystem should respond **no**; but there is no algorithm for determining this in a finite amount of time. To accommodate this situation,

we allow a subsystem to return **und** (undecided) as one of its answers.

Let R be a resource bound. If $\mathbf{M}$ derives a query ϕ within this bound, we write $\mathbf{M}(\phi, R)$:yes; if it decides that ϕ is not derivable, we write $\mathbf{M}(\phi, R)$:no; and if it cannot decide one way or the other within the bounds R, we write $\mathbf{M}(\phi, R)$:und. (We abbreviate $\forall r.\mathbf{M}(\phi, r){:}x$ by $\mathbf{M}(\phi){:}x$.) Note that real agents are computationally oriented; the inference rules specify which derivations are possible, but the subsystem has the option of responding **und** if its resources are not sufficient to actually compute a derivation.

The response of the introspective machine in Figure 1 is extended in the following way: whenever $\mathbf{IM}$ returns **und** on ϕ, $\mathbf{M}$ returns **und** on both $\Box\phi$ and $\neg\Box\phi$. We can summarize the response of $\mathbf{M}$ to self-belief queries of the form $\Box\phi$ and $\neg\Box\phi$ by considering the behavior of the $\mathbf{IM}$ on ϕ. R is the bound for the self-belief query, and R' for the introspective query.

$$
\begin{array}{lcll}
\mathbf{IM}(\phi, R'){:}\text{yes} & \rightarrow & \mathbf{M}(\Box\phi, R){:}\text{yes}, & \mathbf{M}(\neg\Box\phi, R){:}\text{no} \\
\mathbf{IM}(\phi, R'){:}\text{no} & \rightarrow & \mathbf{M}(\Box\phi, R){:}\text{no}, & \mathbf{M}(\neg\Box\phi, R){:}\text{yes} \\
\mathbf{IM}(\phi, R'){:}\text{und} & \rightarrow & \mathbf{M}(\Box\phi, R){:}\text{und}, & \mathbf{M}(\neg\Box\phi, R){:}\text{und}
\end{array}
$$

Note that we want to leave open the possibility that a real agent has no knowledge of some of her own beliefs, and this is where the "undecided" answer plays a crucial role. If $\mathbf{IM}$ returns **und**, then $\mathbf{M}$ will be undecided about its introspective belief.

One obvious result of the imposition of resource bounds is that condition (2) of Definition 1 must be abandoned for sufficiently hard underlying languages. Further, we may also have to give up condition (3). Given resource bounds, the behavior of $\mathbf{IM}$ may differ significantly from $\mathbf{M}$, even when they have the same base sentences and inference rules. For example, let the query to $\mathbf{M}$ be the sentence $\alpha \wedge \Box\beta$ with resource bound R. The control strategy of $\mathbf{M}$ might apply a rule to break this sentence into two conjunctive subqueries, α and $\Box\beta$. The solution of α may consume a large fraction or $\mathbf{M}$'s computational resources. Thus when it asks $\mathbf{IM}$ to solve the query β, it may give $\mathbf{IM}$ a significantly lower resource bound than R. Thus

although **M** would respond yes to β posed *simpliciter*, it won't be able to derive the subquery $\Box\beta$, because **IM** does not have enough resources to do so.

If constraints (2) and (3) of Definition 1 do not hold for real agents, can we find weaker correspondents? For condition (2) we have already done the best we can, by assuming that the answers returned by a subsystem respect the intended semantics of the underlying language and self-belief operator. Because of this, real agents as we have defined them obey a *monotonicity* condition: for $R' > R$, the only difference in the behavior of a belief subsystem can be to change some undecided queries to yes or no. Thus a belief subsystem with a large resource bound is never further away from consequential closure than one with a small bound. However, we may not want real agents to abide by monotonicity — perhaps, if a query cannot be derived within a resource bound, we may want to jump to the conclusion that it cannot be derived.[2] In this sense our definition of belief subsystem may be too strict for some purposes.

We can obtain weaker versions of condition (3) by considering two interesting constraints between **IM** and **M**: *faithfulness* and *fulfillment*. Roughly, an **IM** is faithful if whenever it returns a definite answer (yes or no) on a query, **M** also returns the same answer on that query. Fulfillment is the converse: whenever **M** returns a yes (or no) on a query, **IM** must also.

Definition 3 *An introspective belief subsystem* **M** *is* faithful *if it has the following properties for every introspective pair* $\mathbf{I}^n\mathbf{M}$, $\mathbf{I}^{n+1}\mathbf{M}$:

positive faithfulness (*pfa*):

$$\exists r.\mathbf{I}^{n+1}\mathbf{M}(\phi, r)\text{:yes} \rightarrow \sim \mathbf{I}^n\mathbf{M}(\phi)\text{:no}$$

negative faithfulness (*nfa*):

$$\exists r.\mathbf{I}^{n+1}\mathbf{M}(\phi, r)\text{:no} \rightarrow \sim \mathbf{I}^n\mathbf{M}(\phi)\text{:yes}$$

[2] This type of nonmonotonic reasoning differs from that of McCarthy [10], which is based instead on the notion of a minimal model of a theory. McCarthy is not concerned with the problem of resource-limited derivation, but rather with the inability or undesirability of stating all conditions which do not obtain in a situation.

Proposition 6 *In a faithful introspective belief subsystem* **M**,

pfa: $\exists r.\mathbf{M}(\Box\phi, r)$:yes $\rightarrow$ $\forall r.\sim\mathbf{M}(\phi, r)$:no

nfa: $\exists r.\mathbf{M}(\neg\Box\phi, R)$:yes $\rightarrow$ $\forall r.\sim\mathbf{M}(\phi, r)$:yes

This proposition follows readily from the definition of faithfulness and the monotonicity of responses with increasing resource bounds. Faithfulness is about the weakest constraint we can impose on introspective systems, and is a kind of soundness condition on introspective reasoning. That is, **IM** should not contradict **M**, in the sense that if **IM** ever decides a query, **M** should never decide the opposite.

Definition 4 *An introspective belief subsystem* **M** *is* fulfilled *if it has the following properties for every introspective pair* $\mathbf{I}^n\mathbf{M}$, $\mathbf{I}^{n+1}\mathbf{M}$ *and resource* R:

positive fulfillment (*pfu*):
$$\mathbf{I}^n\mathbf{M}(\phi, R)\text{:yes} \rightarrow \mathbf{I}^{n+1}\mathbf{M}(\phi, R)\text{:yes}$$
negative fulfillment (*nfu*):
$$\mathbf{I}^n\mathbf{M}(\phi, R)\text{:no} \rightarrow \mathbf{I}^{n+1}\mathbf{M}(\phi, R)\text{:no}$$

Proposition 7 *In a fulfilled introspective belief subsystem* **M**,

pfa: $\mathbf{M}(\phi, R)$:yes $\rightarrow$ $\exists r \geq R.\mathbf{M}(\Box\phi, r)$:yes

nfa: $\mathbf{M}(\phi, R)$:no $\rightarrow$ $\exists r \geq R.\mathbf{M}(\neg\Box\phi, R')$:yes

This proposition follows from the definition of fulfillment and the monotonicity of definite responses. Fulfillment is a kind of completeness property for introspection, in the sense that, if **M** derives ϕ, there is some resource bound at which it will also derive $\Box\phi$ (or $\neg\Box\phi$, if ϕ is not a belief).

Fulfillment and faithfulness are not independent. For example, if we take the contrapositive of *pfu*, we get (for arbitrary R) $\sim\mathbf{IM}(\phi, R)$:yes $\rightarrow$ $\sim\mathbf{M}(\phi, R)$:yes. But if **IM** responds no to a query ϕ, it never responds yes to the same query, so we also have $\exists r.\mathbf{IM}(\phi, r)$:no $\rightarrow$ $\sim\mathbf{IM}(\phi, R)$:yes.

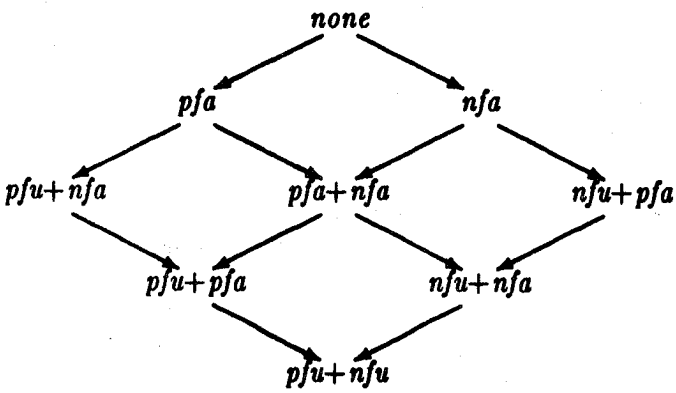


Figure 2: Lattice of domination relations

By the elementary laws of propositional logic, we then can derive $\exists r.\mathbf{IM}(\phi, r)$:no $\rightarrow \sim\mathbf{M}(\phi, R)$:yes from the contrapositive of *pfu*, and this is the condition *nfa*, because R is arbitrary. In a similar manner, we can show that negative fulfillment implies positive faithfulness ($nfu \rightarrow pfa$).

These two relations are the only interdependencies of the constraints. There are thus nine distinct combinations that can be arranged in a lattice as in Figure 2.2. The arrows indicate domination relations; the constraint $pfu+nfu$ is thus the strongest of the possible conditions on introspective belief, in the sense that every introspective belief subsystem that obeys it also obeys every other possible combination of the faithfulness and fulfillment constraints. Note that positive fulfilled systems dominate negative faithful ones, and negative fulfilled systems dominate positive faithful ones.

Example 1 The use of introspective belief subsystems as a descriptive model of human belief will be illustrated with one example, drawn from Hintikka [3]. He argues that if someone believes ϕ, she also believes that she believes it (at least in the absence of strict resource limitations on reasoning). This is our condition of positive fulfillment, where the resource R is always taken to be arbitrarily large, and we consider only the first level of introspection (**M** and **IM**). Hintikka goes on to argue that people will often have

false ideas about their own beliefs, e.g., an utterance of the form

> *S believes that she believes that ϕ although she does not believe it* (1)

can be a true statement about the state of S's beliefs.[3] In terms of the introspective model, we would say that human belief subsystems are not positive faithful (and hence not negative fulfilled).

There is an additional curiousity to Hintikka's theory. Although the first level of introspection is characterized as being positive fulfilled but not necessarily positive faithful, it appears that subsequent levels are considered to be totally faithful. For example, the utterance

> *S believes the following: that she believes that she believes ϕ, although she does not believe it* (2)

which is the statement of (1) as applied to S's idea of herself, is taken to be always false. In our introspective model, this is a statement about self-belief sentences of the introspective machine **IM**. To capture this behavior, we simply let **IM**'s concept of self-belief be positive faithful.

2.3 Computational Issues

We now present some of our computational results on introspective machines. Generally, we are interested in the problem of converting a nonintrospective belief subsystem into an introspective one; one can imagine retrofitting an existing knowledge base with a mechanism for reasoning about its own beliefs. The questions we pose will have the following form: given a particular introspective constraint (a point in the lattice of Figure 2.2), and perhaps other conditions on nonintrospective behavior, can we implement a belief subsystem obeying these constraints? That is, we would like to find an algorithm that will return a definite answer (yes or no) to every query, given the constraints, so that the introspective belief subsystem is de-

[3] This is sentence 83 on page 125 of Hintikka [3].

cidable. We first make this notion of decidability precise for resource-limited agents.

Definition 5 *Let $R(\phi)$ be a function mapping queries into finite resource bounds. A belief sybsystem* **M** *is decidable if there exists an algorithm and function R for* **M** *such that for all queries ϕ, ~**M**(ϕ, $R(\phi)$):und; it is semi-decidable iff whenever ϕ is derivable,* **M**(ϕ, $R(\phi)$):yes; *it is undecidable if it is not semidecidable.*

The following proposition relates real and ideal agents.

Proposition 8 *Suppose a real introspective belief subsystem* **M** *obeys the following constraints:*

1. **M** *is consistent.*
2. *The inference rules of* $\mathbf{I}^n\mathbf{M}$ *are complete for all* $n \geq 0$.
3. **M** *is fulfilled (*pfu *and* nfu *hold).*

Then **M** *is an ideal introspective belief subsystem iff it is decidable.*

The proof is to show that all three conditions of an ideal introspective agent in Definition 1 are satisfied. The first two obviously are. By inspection, we note that the constraint *pfu+nfu* means that all $\mathbf{I}^n\mathbf{M}$ have the same behavior; hence the third condition is satisfied. Finally, if **M** is decidable, there is a function $R(\phi)$ for which **M** always returns a definite answer; hence the belief set of **M** is the same as that of an ideal agent. Note that a real agent is ideal only if she has an algorithm that will decide any query ϕ in the finite resource bound $R(\phi)$. Real agents are always computational.

Now let us assume the first two conditions of Definition 8 hold, and explore the computational nature of belief systems obeying various introspection conditions. By "nondoxastic **M**" we mean that the base set of every belief subsystem of **M** is nondoxastic.

Proposition 9 *Let the introspection constraint be* pfu+nfu. *If the underlying language is*

1. *semidecidable,* **M** *is undecidable;*
2. *propositional, nondoxastic* **M** *is decidable;*
3. *monadic PC, nondoxastic* **M** *is undecidable.*

This proposition just collects the results of the last section (Propositions 3–5 with respect to real agents. Note that, except in the case of a propositional language, **M** must return und for some queries, no matter what resources are available. In these cases, real agents are not even approximations of ideal agents, since there is no limit in which their behavior becomes the same).

Now suppose we are given a nonintrospective belief subsystem **M**(whose base set is nondoxastic), and we are asked to construct an introspective subsystem **M'** whose first component is **M**. We are free to choose the introspective components, as long as they satisfy conditions (1) and (2) of Proposition 8. The following proposition tells us the best we can do in terms of satisfying various introspective constraints.

Proposition 10 *Suppose the underlying language of* **M** *is decidable. Then if the introspection constraint is*

1. pfu+nfu, **M'** *is undecidable;*
2. pfu+pfa, **M'** *can be semidecidable;*
3. nfa+pfa, **M'** *can be decidable.*

The first result is simply (1) of Proposition 9. The second says that if we only want to enforce positive fulfillment and positive faithfulness, the best we can do is to construct an introspective subsystem that is semidecidable. And finally, if the introspection constraint is simple faithfulness, we can construct a decidable **M'**. Of course, we can do better than this for particular underlying languages (*e.g.*, propositional), but there exists a decidable language for which these bounds are strict (namely, monadic PC).

Let us put these results into perspective. If we are given a nonintrospective agent whose inference rules are com-

plete and whose beliefs are decidable, the best we can do in retrofitting introspective reasoning is to make the agent's self-beliefs faithful. However, if we start with an agent whose rules are incomplete, or we are willing to give up completeness, we can enforce stricter introspective constraints. But now these constraints are relative to a much weaker notion of belief derivation. For example, suppose an agent has no inference rules at all, so that her only nonintrospective beliefs are the base sentences. Certainly we can form a decidable introspective belief subsystem in which *pfu+nfu* holds; $\Box\phi$ is a belief if ϕ is a member of the base sentences, and $\neg\Box\phi$ is a belief if not, and membership in the finite base set is decidable.

3 Comparison to Related Work

Our definition of an ideal introspective agent has many points of similarity with work by Halpern and Moses [2] and Moore [12]. In both these latter cases an underlying propositional language is used, and beliefs sets are defined nonconstructively as *stable sets* (Stalnaker [13], although his original definition did not include consistency).

Definition 6 *A* stable set S *obeys the following constraints:*

1. S *is consistent.*
2. S *is closed under truth-functional consequence*
3. *If* $\phi \in S$, *then* $\Box\phi \in S$; *and*
   if $\phi \notin S$, *then* $\neg\Box\phi \in S$.

Now we would like an ideal rational agent's beliefs to be a stable set. To build an agent with ideally rational beliefs, we require favorable answers to the following questions.

(a) Given a sentence α that represents the initial beliefs of an agent, what is the appropriate stable set containing α that should be the belief set of the agent?

(b) Is there an algorithm for computing it?

The answer to (a) is not as simple as might be supposed, because it involves finding a stable set that includes α, and makes the fewest assumptions about what the agent believes in addition to α. The presence of doxastic sentences in α complicates matters, and indeed Halpern and Moses differ from Moore in identifying an appropriate belief set. However, if α is consistent and nondoxastic, both approaches converge on a single stable set. Further, this stable set is identical to the belief set of an ideal introspective agent with base set α, so that by Proposition 4 there exists an algorithm for deciding membership in the stable set (the algorithm D^{α} of Halpern and Moses [2] decides the stable set in this case).

From Definition 1 and Propostion 2, and ideal introspective subsystem, if it exists, is a stable set. Further, for the propositional case it yields the appropriate stable set, in the sense of question (a) above, taking α to be the base set of **M**. Now we can use the results of Section 2.1 to analyze the computational nature of stable sets in the case of quantified languages. By proposition 5, even for the relatively simple case of monadic PC and nondoxastic α, the question of membership in the stable set is undecidable. Thus for these systems we must answer question (b) in the negative.

4 Conclusion

We have developed a theory of introspection based on the idea that an agent can use a model of her own belief subsystem to reason about self-belief. The theory can serve as a descriptive tool, since we can describe agents with varying degrees of self-knowledge; hence it may be useful to researchers interested in modelling the cognitive state of users (*e.g.*, in domains such as natural-language systems, tutoring systems, intelligent front ends to databases, and so on). The theory also is a guide to building agents with introspective capabilities, or retrofitting these capabilities onto existing artificial agents.

Introspective belief subsystems can be related to the standard propositional modal logics for belief, weak $S4$

and $S5$ (the axiom schema $\Box p \supset p$ is discarded). An ideal introspective agent is described by weak $S5$ plus a consistency axiom $\Box p \supset \neg\Box\neg p$, since by Proposition 2 both $\Box p \supset \Box\Box p$ and $\neg\Box p \supset \Box\neg\Box p$ are true of such agents. An introspective agent with complete inference rules obeying *pfu* is described by weak $S4$ plus consistency. There are no standard epistemic logics for agents which simply obey the faithfulness constraint; we could construct these by adding $\Box\Box p \supset \Box p$ and $\Box\neg\Box p \supset \neg\Box p$ to the modal logic K.

There are many interesting questions about introspective subsystems that have not been answered in this paper, especially relating to ideal agents. There is obviously a close connection between our definition of an ideal introspective agent and the autoepistemic theories of Moore [12], yet we have compared them only for the case of nondoxastic base sets. Given a (perhaps doxastic) sentence α, Moore defines T to be a ***stable expansion*** of α if T is equal to the set of truth-functional consequences of

$$\{\alpha\} \bigcup \{\Box p : p \in T\} \bigcup \{\neg\Box p : p \notin T\}.$$

Some sentences have no stable expansions, some have just one, and some have more than one. For example, $\alpha = (\neg\Box P \supset Q) \wedge (\neg\Box Q \supset P)$ has two stable expansions, one containing P, the other Q. What happens to an ideal introspective subsystem when α is its base set? Given the query P, **M** will try to prove $\neg\Box Q$, and issue the query Q to **IM**. **IM** will then try to prove $\neg\Box P$, and issue the query P to $\mathbf{I^2M}$. Thus there is no terminating derivation of P, and similarly for Q. However, at some point we could notice that the query delivered to $\mathbf{I^nM}$ is exactly the same as that for $\mathbf{I^{n-2}M}$, and decide that there is no derivation of the query. If we decide this when the query is P, we will get the stable set containing Q; conversely, if we decide that Q is not derivable, we will arrive at the stable set containing P.

Although this example is suggestive, we do not yet have any definitive results on the relationship between Moore's autoepistemic theories and ideal introspective subsystems.

References

[1] Collins, A. M., Warnock, E., Aiello, N. and Miller, M. (1975) Reasoning from Incomplete Knowledge. In *Representation and Understanding*, Bobrow, D. G., and Collins, A., eds., Academic Press, New York.

[2] Halpern, J. Y. and Moses, Y. (1984). Towards a Theory of Knowledge and Ignorance: Preliminary Report. Computer Science Research Report RJ 4448, IBM Research Laboratory, San Jose, California.

[3] Hintikka, J. (1962). *Knowledge and Belief.* Cornell University Press, Ithaca, New York.

[4] Konolige, K. (1984). Belief and Incompleteness. Artificial Intelligence Center Technical Note 319, SRI International, Menlo Park, California.

[5] Konolige, K. (1984). *A Deduction Model of Belief and its Logics.* Doctoral thesis, Stanford University Computer Science Department Stanford, California.

[6] Kripke, S. A. (1959). A Completeness Theorem in Modal Logic. *Journal of Symbolic Logic* **24**, pp. 1–14.

[7] Kripke, S. A. (1962). The Undecidability of Monadic Modal Quantification Theory. *Zeitschrift für Mathematische Logik and Grundlagen der Mathematik* **8**, pp. 113–116.

[8] Kripke, S. A. (1963). Semantical considerations on modal logics. *Acta Philsophica Fennica* **16**, pp. 83–94.

[9] Levesque, H. J. (1982). A Formal Treatment of Incomplete Knowledge Bases. FLAIR Technical Report No. 614, Fairchild, Palo Alto, California.

[10] McCarthy, J. (1980). Circumscription — A Form of Nonmonotonic Reasoning. *Artificial Intelligence* **13**, pp. 27–39.

[11] Moore, R. C. (1980). Reasoning About Knowledge and Action. Artificial Intelligence Center Technical Note 191, SRI International, Menlo Park, California.

[12] Moore, R. C. (1983). Semantical Considerations on Nonmonotonic Logic. Artificial Intelligence Center Technical Note 284, SRI International, Menlo Park, California.

[13] Stalnaker, R. (1980). A Note on Non-monotonic Modal Logic. Unpublished manuscript, Department of Philosophy, Cornell University.

Languages with Self-Reference I: Foundations

(or: We Can Have Everything in First-Order Logic!)

Donald Perlis
University of Maryland, College Park, MD 20742, U.S.A.

Recommended by Nils Nilsson

ABSTRACT

It is argued that a proper treatment of cognitive notions such as beliefs and concepts should allow broad and consistent expression of syntax and semantics, and that this in turn depends on self-reference. A theory of quotation and unquotation is presented to this end that appears to make unnecessary the usual hierarchical and non-first-order constructions for these notions. In the current paper (Part I) the underlying theory is presented; a sequel will treat in more detail the applications to cognition.

1. Introduction

Language provides a kind of discrete representation of reality. This serves the purpose of facilitating planning by providing discrete computable representations (statements) of what is possible in a given domain. So viewed, statements are part of the very world they purport to describe. It is our contention that a proper understanding of language will in the long run be found to depend upon this kind of self-referential ability. Herein is found the main theme of this paper: the interplay of assertion and meaning, syntax and truth.

All syntactic features then should be themselves expressible in quoted form, allowing the user to inspect and comment on usage. "The user" here means the reasoning system, be it a human, robot, or other. The kind of flexible system envisaged is one that will inevitably make errors due to the complexity of its environment, and that will then have on occasion to re-evaluate its representations. This is the reason that we insist on a language that has expressions naming its own expressions: it will be crucial to be able to isolate a statement as such, and state that it is in error, and even to point out the exact expressions that should be changed.

Artificial Intelligence **25** (1985) 301–322
0004-3702/85/$3.30

It is worth dwelling on this further. Some proposals (see, e.g., [1]) simply view old assertions to have been taken away or deleted, rather than noting them to be imperfect. That is, the old assertion is simply gone altogether by definition in the new state of the reasoning mechanism. As an example, the conclusion that Tweety can fly, based on the information that Tweety is a bird and on the absence of information that she cannot fly, disappears on the introduction of new information to the effect that Tweety cannot fly, with no trace of the former conclusion to the contrary. But then the fact of an error having been made and then corrected is not itself represented and so is not something that can be reasoned about with these same mechanisms. This leads to the usual complaint that process information is not cleanly represented in the system itself. We seek here to devise a language closer in spirit to natural language in that the language can be reasoned about in that same language itself.

As a consequence, a principal concern of this paper is the need to refer to certain statements as true (and false). The notion of truth appears a crucial one for any treatment of information that is utilized inside a reasoning system whose own behavior is itself to be reasoned about. This is easy to see: even though we may not *know* a statement of ours to be true, we do want to be able on certain occasions to consider that it may be so (or that it may not). Without such a capability, error correcting, even of the deletion sort, cannot be done. What is being urged here, is to let such judgements come out in the open, so that the reasoning apparatus can be brought to bear on them.

As an example of the kind of reasoning we have in mind, consider a reasoner R that has belief B, then later learns not-B, recalls having believed B, and concludes that the statement that all its beliefs are true was itself false and therefore may well still be false. This involves temporal and inductive inference as well as a number of other things. We do not address all the aspects of this problem here; it does however illustrate the importance of the roles of truth and self-reference in commonsense reasoning. (See [2] for a broader discussion of the various aspects of the kind of reasoning just mentioned.)

In this Part I of our investigation, we concentrate attention on the first-order truth-definitional issues; more detailed analyses of cognitive notions and difficulties with modal and other treatments in the literature will be given in a sequel. Suffice it to say here that these approaches do not allow for unrestricted reference to syntax, and so do not satisfy our expressive criteria.

2. Some History

Let us start by reviewing some history. Gottlieb Frege developed the first formal quantificational logic over a period of more than two decades culminating in 1903. This consisted of a precise syntax, a set of inference rules, and axioms, where only one kind of variable and constant was employed. The idea

was to have a *universal* language for logic, in which no a priori distinction was granted between primitive individuals and fancier constructs involving those primitive individuals. Thus for Frege, an object c and the properties P it may have were all objects to be reasoned about in the same way, i.e., with the same basic rules and notations. Frege had certain *comprehension axioms* that specifically created object-notations "P" for properties P, and stated that sentences using properties as predicates could be equivalently reprased using properties as objects. These axioms in effect state a relationship between a name and what it names:

$$\text{Has}(c, \text{“}P\text{”}) \leftrightarrow P(c)\,,$$

or equivalently, but closer to Frege's notation:

$$c \in \{x \mid P(x)\} \leftrightarrow P(c)\,.$$

In that same year Bertrand Russell showed that Frege's system was inconsistent. (Specifically, he defined the property $R(x)$ so that $R(x) \leftrightarrow -\text{Has}(x, x)$, and applied this to $x = \text{“}R\text{”}$.) Russell then proposed that objects be arranged in a hierarchy with different notations and rules, thus avoiding the possibility of self-reference that led to the inconsistency in Frege's system. The resulting "typed" system has as its first level of notations precisely that of Frege, but without the damaging axioms (the comprehension axioms) that created objects out of properties at the first level. For Russell, properties of first-level objects are to be viewed as second-level objects. Thus was born the expression "first-order logic"; it is Frege's logic except for the offending axioms. Russell's logic included this first level and also higher levels for objects corresponding to properties, properties of properties, etc. (which for Frege would have been created by comprehension axioms at this first universal level). For this reason we refer to first-order logic and the higher-order logics. In general one can choose to work with as many orders as desired. Names for objects are supplied at each level by the rules of syntactic formation.

However, some problems do arise. One is that there is a substantial burden in having to deal with a large number of different notations. This perhaps could be excused, although the apparent fact that in natural languages such as English we have no need for levels may suggest that a better approach exists. A second problem is more serious: many significant concepts cannot be expressed at all with levels, as will be seen below. The original simplicity and plausibility of Frege's approach has then continued to attract interest, and much of modern logic has been motivated by efforts to revise it to preserve its desirable features while removing inconsistency. Artificial intelligence has come to join this effort, as it became recognized that more than Russell's higher-order individuals are required in many situations. We give two examples here; others will appear as we proceed.

(i) *Universal quantification.* Suppose that agent A's beliefs are represented as sentences in some formal language L with levels. Then symbols in L are indexed by their levels, e.g., ti for a constant or variable or function of level i, and $P/i + 1/(ti)$ for a predicate of level $i + 1$ applied to a term of level i. Then the sentence that A has no religious beliefs, we might try to formalize as

$$(xi)(\text{Bel}/i + 1/(A1, xi) \rightarrow \text{Not-religious-belief}/i + 1/(xi)) .$$

But this is not quite what the original sentence says, for we need not think any supposed religious beliefs of A to be at any particular level i. There is no way to quantify over all levels at once and stay within the framework of levels at the same time. On the other hand, we *want* to write simply

$$(x)(\text{Bel}(A, x) \rightarrow \text{Not-religious-belief}(x)) .$$

(ii) *Existential quantification.* Consider the sentence that John has a false belief. Again we might write $(Exi)(\text{Bel}/i + 1(\text{John}, xi)$ & $\text{False}/i + 1(xi))$. But we don't know what level John's supposed false belief is at, so really we'd need to write something like $(Ei)(Exi)(\ldots)$, i.e., for some level i John has a false belief xi. But then i is being used as a variable and so needs a level of its own, in opposition to our intention of using any level at all as a possible substitute for it. Again, we would like to write simply $(Ex)(\text{Bel}(\text{John}, x)$ & $\text{False}(x))$.

To be sure, an infinite set of hierarchical sentences will do the trick in the first example: one for each value of the variable xi and index i. For instance, this is allowable in Konolige [3]. But this doesn't provide then a (single) expression that can be reasoned with. No one could ask the system the question as to whether the given assertion is the case; it would take forever to ask! Moreover, to deny the assertion would involve a single infinitary disjunction, which is also what happens in the second example: either John has a level-1 false belief, or a level-2 false belief, or etc.

What seems to be needed is an avoidance of separate levels altogether, so that all concepts are treated at the same (first) level. For instance, McCarthy [4] has usefully introduced names for concepts (second-order objects) into a first-order system. McCarthy considers the problem of distinguishing between a phone number as a number, and a phone number as a concept (that which is dialed on a phone to reach so-and-so). This is of importance, since we don't want to say Alice knows Bill's phone number simply on the basis that she knows of the number 2345679, e.g., it may be her bank account number. Still, this may indeed be his number, and if we write Bill's-number = 2345679 then we are in trouble, for she knows the latter but not the former, or so we want to say. (Here "knows" can be taken to mean "has in mind" or "has memorized".)

McCarthy shows that this issue can be resolved by viewing Bill's-number as something different from the number itself, namely, as a second-order con-

struct which however is embeddable in a first-order setting as a new kind of individual: a concept. That Bill's-number can be viewed as a second-order construct is seen as follows: it is a relation between Bill and a number, expressible as Phone-Number(Bill, 2345679). Thus Alice may have memorized 2345679, but not the fact that this is Bill's number. The latter situation could be expressed as

Knows(Alice, "Phone-Number(Bill, 2345679)") ,

and even more usefully as

Ex[Digits(x) & Knows(Alice,concat(x," = number(Bill)"))] .

I.e., Alice knows Bill to be related to a particular string of digits, in a "Phone-Number" sort of way. This allows for the expression of such a notion as Alice knowing that Bill's number (concept thereof) is not to be found in the phone book: Knows(Alice("$-(Ex)$(Phone-Number(Bill, x) & Listed (x))")).

Note that the Digits(x) is important, to prevent using x = "number(Bill)". Here Alice *knows* Bill's phone number in the sense of knowing that such-and-such digits are his phone number, rather than knowing *of* his number. She may have heard many numbers in connection with Bill: his mother's number, his work number, his house number; she knows *of* them all; but it is his home phone number that she knows to be just that. So there is indeed an implicit sentence that she knows (believes). What Alice knows is not a number, but a fact relating Bill and a number. For this, quotation, in one form or another, is needed. See Haas [5] for another notational version, and for more detail on algorithms for reasoning in this vein. The *concat* function is a method for introducing variables into string expressions, a form of *quasi-quotation* or *quantifying in.* Concat ("a", "b") is interpreted as "ab", whereas concat(x, "b") remains uninterpreted until x is assigned a meaning. On the other hand, concat("x" "b") is "xb", so that we have flexibility in our use of variables: opaque or transparent as desired.

McCarthy suggests that introducing function symbols for concepts (rather than quoted expressions) may be sufficient for a general treatment of concepts. However, it seems to us that this hope is unfounded. We often depend on expressions in the formation of concepts. For example, the concept of a sentence derives its usefulness from being related to particular sentences and particular words that make up sentences: "the last word you just said" is an expression which although representable as a function still refers to a particular word, not to a concept. Thus quotation seems necessarily involved at some point if we are to have a self-describing language. It appears we must describe specific expressions as carriers of (the meanings of) concepts. In any case, even the strict use of functions for general concepts will lead to paradoxical

situations unless care is taken. In [5] a functional approach is taken that otherwise is along the lines suggested here.

Thus it appears reasonable to allow a certain number of levels to be 'collapsed' into first-order logic, and leave the rest out (either entirely out or represented in Russell's higher types). Now, the question arises, why not collapse all levels into first-order logic, and be done with these difficulties? This however is just what causes Russell's paradox. McCarthy and others—Elschlager [6], Weyhrauch [7], Attardi and Simi [8]—are careful to avoid contradiction, by not using full comprehension axioms, and indeed no need for them arises in limited cases such as these. But if we wish to address the issues raised in the introduction, and in particular the two sample sentences (i) and (ii), then we must find a way to collapse all levels into one without contradiction, i.e., we need to have a self-referential or universal language. (Indeed, Creary [9] makes some effort to carry out such a program, along somewhat hierarchical lines, and observes the need to address possible inconsistencies.)

3. Names

Because of these difficulties the approach of levels appears too restrictive for artificial intelligence in general. Montague [10] argues that modal logic is the appropriate remedy, and that this yields a consistent treatment of epistemic notions, whereas first-order logic (FOL) with the same notions, is not. But this is due to a powerful strengthening of FOL in writing variables that can be replaced by names for formulas; if his modal logic is similarly endowed, i.e., with propositional variables, then it too appears inconsistent. We will investigate this in formal detail elsewhere. See also Burge [11], for another criticism of Montague's position.

To motivate another approach, let us consider an extended example. Consider the English sentence, "John believes that Ronald Reagan is President." This we could formalize as

S1: Bel(John, "President(Ronald Reagan)")

where "Ronald Reagan" is a constant term, and our formal language has constant names for sentences (here the second argument to Bel names the sentence inside quotes). Here we have adopted the ideas of Moore and Hendrix [12], regarding beliefs as forming a set of sentences.

Now, it is essential to have also an un-naming device that would return a quoted sentence to its original (assertive) form, together with axioms stating that that is what naming and un-naming accomplish. This would make it possible, for example, for another agent, say Sally, to reason from the sentence S1 above that John has a true belief (assuming she also believes Reagan is President) or that John will answer "Ronald Reagan" when asked who is

President. She could put herself in John's shoes, un-naming his beliefs, and then reason with the results, being careful to avoid using other beliefs of her own that she feels are not also ones of John's. (Haas [5] and Konolige [3] study aspects of this.)

Consider the further sentence

S2: "There is someone whom John believes to be President."

This we might try to formalize as (Ex)Bel(John, "President(x)"). Something is wrong here. The inner "x" is not recognized by the syntax of FOL; there's simply a single constant "President(x)". What we want is a way to take an arbitrary x and form from it a sentence name "President(x)". So let's do just that: let concat be a new 4-place function symbol, where concat(a, b, c, d) intuitively stands for a name of the concatenation of whatever a, b, c and d are names of. (If some argument isn't a name, concat can default to some convenient constant.) Then we can write

$$(Ex)(\text{Bel}(\text{John}, \text{concat}(\text{“President”}, \text{“(”}, x, \text{“)”}))) .$$

Now here an appropriate witness to John's existential belief—i.e., an object filling the role of the x required to exist—is not Ronald Reagan but rather "Ronald Reagan". For only the string naming Reagan concatenates with "President" etc., to give the desired name "President(Ronald Reagan)". This may appear clumsy and even counter to the sense of S2; but note that unless there is a description that John can use to refer to Reagan, it makes little sense to say that he has the belief in question. For the squeamish, we can be a little fancier:

$$(Ey)(Ex)[\text{Names}(\text{John}, x, y) \\ \& \text{Bel}(\text{John}, \text{concat}(\text{“President”}, \text{“(”}, \text{“)”}))]$$

i.e., there is a person y and an object x (that John uses as a name for y) for which John believes the indicated sentence (namely, in this case, "President(Ronald Reagan)"). Here y is Reagan himself, and x is "Ronald Reagan". Note that the second extended version above actually entails the first.

This provides a solution to a problem pointed out by Moore [13]. For instance, Moore mentions the following rule we may want to adopt for the predicate Knows:

$$\text{Knows}(a, \text{“}p \rightarrow q\text{”}) \ \& \ \text{Knows}(a, \text{“}p\text{”}) \rightarrow \text{Knows}(a, \text{“}q\text{”})$$

where "a" stands for a person. If we want to be able to use such a rule for

arbitrary p and q we must use variables in place of p and q. If we quote the variables, this could mean inventing special string matchers, as Moore warns. But using concat, it is fairly direct:

$$\text{Knows}(a, \text{concat}(x, \text{“}\rightarrow\text{”}, y))$$
$$\& \ \text{Knows}(a, x) \rightarrow \text{Knows}(a, y)\,.$$

(In fact Moore does something like this, although more complicated, along with his possible-worlds treatment; however we need not follow him that far to get what we need. The decision to represent knowledge as quoted sentences, together with variables ranging over such sentences, already holds enough for us.)

Here we have used concat with only three arguments, so we had best tell the whole story about it and about names: we really want concat to be a 2-place function symbol, and when we write concat($a, b, \ldots, n$) we are abbreviating concat(a, concat(b, (concat($\ldots$ (m, n))) $\ldots$)). There are many ways to create names. One that is both simple and general is as follows: First we employ a form of Hollerith quotation, i.e., $n{:}a1 \ldots an$ is a name for the string $a1 \ldots an$ of the n symbols ai. These names are new "compound" constant symbols, not counted as single symbols when forming a name in which they appear. Thus a name for $(x = y)$ is $5{:}(x = y)$, and a name for *this* is $7{:}5{:}(x = y)$ rather than $1{:}5{:}(x = y)$ even though $5{:}(x = y)$ is a single compound symbol as well as a string of seven simple symbols. Note that the colon is counted, and its name is $1{::}$. This form of quotation is used to avoid the problem of nested quotation marks; now we have ability to name arbitrary strings made of *any* symbols in our language including those used in the quotation mechanism itself. Then we require

$$\text{concat}(n{:}a1 \ldots an, m{:}b1 \ldots bm) = n + m{:}a1 \ldots bm\,.$$

We note that in Haas [5] an alternative notation is used that has some simplifying advantages, although less generality. In the sequel we will however revert to quotation marks for the most part, it being understood that the Hollerith form is available to sort out ambiguities.

Now that we have motivated the need for an un-naming or un-quoting device, let us see whether it can be obtained in a form that is general, useful (practical), and consistent. Prior results by Tarski [14] and Montague [10] suggest that our goal may be unobtainable. However, logicians have continued to explore ways of capturing the intuitive sense of Frege's system without the inconsistencies, and we shall exploit and combine some of this work to achieve an apparently satisfactory treatment.

4. A New Approach to Truth

Names do present a difficulty however, namely that found by Russell for Frege's system. In its barest form, it amounts to Tarski's "No Truth-Definition Theorem" [14]: In general,

$$\text{True}(\text{“}A\text{”}) \leftrightarrow A$$

is inconsistent, i.e., unquotation doesn't fully undo quotation. Now if we assume $\text{True}(\text{“}A\text{”}) \leftrightarrow A$, then for certain cases of A—e.g., the famous Liar sentence L: $-\text{True}(\text{“}L\text{”})$—we get

$$\text{True}(\text{“}L\text{”}) \leftrightarrow L \leftrightarrow -\text{True}(\text{“}L\text{”}).$$

Indeed, from identifying $\text{True}(\text{“}P(c)\text{”})$ and $\text{Has}(c, \text{“}P\text{”})$ we would get either of Russell's or Tarski's results from the other.

We must then decide how to eliminate such cases and yet allow benign and useful cases of self-reference. Our original goal of keeping all syntax available for inspection prevents us from simply outlawing certain expressions from the language.

Kripke [15] introduced a brave attack on this classical problem of truth definitions. In order to avoid the consequences of Tarski's theorem to the effect that $\text{True}(\text{“}A\text{”}) \leftrightarrow A$ is in general inconsistent, he suggests that for some formulas A neither $\text{True}(\text{“}A\text{”})$ nor $\text{False}(\text{“}A\text{”})$ hold. This means excluded middle, in the form $\text{True}(\text{“}P\text{”})$ or $\text{False}(\text{“}P\text{”})$, does not hold for all P in Kripke's system. While this can be regarded as a negative feature, leaving 'gaps' in the truth definition, otherwise it has very intuitive behavior.

Tarski's theorem (or the Liar paradox) can be equally regarded as a variant on Russell's paradox, as we will see below. Tarski's own approach to the problem raised by his theorem was similar to Russell's: to introduce a hierarchy of truth predicates, each to apply to formulas formed at stages prior to it. This has the defect of not allowing reference to general formulas; for example $\text{True}(\text{“}A\text{”}) \rightarrow A$, although valid for each truth predicate True when applied to any formula A formed prior to the introduction of that predicate, cannot be stated in one formula for all levels of the hierarchy at once. And the statement that a particular formula B is "not true" at *any* level ($-\text{True}(\text{“}B\text{”})$) is not representable either.

For this reason Kripke introduced his approach using truth gaps, in which there is only one truth predicate. Yet the problems persist on close inspection. Kripke himself comments on the problematic character of gaps:

> "... Liar sentences are not true in the object language ... but we are precluded from saying this in the object language by our

> interpretation of negation and the truth predicate ... The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us."

In effect, we note that for some formulas A, True("A") never appears in Kripke's construction, so that we conclude for ourselves that $-$True("A"), yet this latter formula is accorded no recognition in Kripke's formal apparatus. It may be that False("A") appears, in which case there is no reason for concern; but also A may be paradoxical, such as a Liar sentence, and then neither True("A") nor False("A") will appear, and the formalism doesn't record what to us is salient, namely the very fact that neither of these appeared.

In developing his method, Kripke utilized a procedure for assigning to True (and False) more and more terms until a fixed point is reached. This choice was a conscious departure from the standard (Tarskian) semantics for first-order logic. In Tarskian semantics, the "truth" of atomic formulas in some domain is determined by some external means, and then all the rest follows, including the holding of $-A$ when A is not determined to hold. Thus Kripke's dilemma does not occur in Tarskian semantics, although other problems do seem to, as noted.

Here we want to suggest that in fact Kripke's work provides the basis for a first-order (excluded middle) treatment of truth after all, in which True is an ordinary first-order predicate symbol and "truth" in the sense of holding in models is kept to be the strictly Tarskian one. We begin by noting a simplifying characterization of Kripke's construction.

We suppose a first-order theory T to have a monadic predicate symbol True. We will interpret True(x) intuitively as meaning that the formula named by x is determined to be "true" in Kripke's sense (in some domain). Thus True(x) in particular means that x is grounded (it can be reduced to formulas not involving "True") and the result holds. Now, this is quite a complex notion, and formalizing it straightforwardly by mimicking Kripke's metalanguage construction would require in the object language a rather massive set of axioms including set theory. However, we can do it much more easily simply by formalizing the single iterative step in the construction, namely, deciding the case of True(x) by reducing it to the case of x. Kripke proceeds by saying True("A") if A has already been determined at a previous stage.

The trick follows Gilmore [16]. We simply posit

$$\text{True}(\text{“}A\text{”}) \leftrightarrow (A)*$$

instead of the earlier-mentioned schema (without the star) where the $*$-operator replaces each connective occurrence of the form $-$True("...") in A by True("$-$(...)"). (Here we have to be careful first to pass negation through to predicate letters, and also to rewrite conditionals $A \rightarrow B$ as $B \vee -A$.)

Note that, in general, True("$-A$") is NOT equivalent to $-A$, and thus (as it

turns out) contradictions do not arise from the famous paradoxes. Yet, we still gain the advantages of self-reference because the two expressions are equivalent in case A itself does not contain the predicate "True"—or more generally if A is *positive* in a sense that will be defined later.

The paradigmatic case is the following:

$$\text{True}(\text{“}-\text{True}(\text{‘}B\text{’})\text{”}) \leftrightarrow \text{True}(\text{“}-B\text{”}) .$$

In Kripke's terms, we decide True("−True('B')") only if we already have decided −True("B"), which for Kripke can only be in the form False ("B") and which we here write as True("$-B$") to keep the number of new predicates down. We see then how we can preserve the important property of excluded middle $(A \vee -A)$ for all formulas A. With Kripke, we do not require True("A") ∨ True("$-A$") for all A, but since now $-A$ is not equivalent to True("$-A$"), we avoid a contradiction. True("$-A$") has a stronger meaning than simply $-A$: the former means "$-A$" was found to come out of Kripke's iterative definition, whereas the latter means simply not whatever A says. For instance, if A is True("B") then $-A$ simply attests to the fact that B doesn't come out of Kripke's iterations (while A says that it does); but True("$\dot{-}A$") says $(-A)*$, and this is True("$\dot{-}B$"), i.e., that $-B$ does come out of the iterations.

In most cases, the new schema True("A") $\leftrightarrow A*$ reduces immediately to the former schema True("A") $\leftrightarrow A$, and when −True does appear in A, the rule is still trivial to apply and intuitively sensible. For example, True("−True('1 = 2')") is equivalent according to the above schema to (−True("1 = 2"))*, which is just True("−(1 = 2)"). This in turn is equivalent to −(1 = 2). We remind the reader that a slightly different quotation mechanism is preferrable, namely Hollerith quotation as before, since then there is no scoping ambiguity, but this isn't necessary for purposes of illustration here.

Thus whenever "True" appears as a predicate letter in $A*$, it will not be negated, and so nothing is ever asserted to be ungrounded in $A*$. The schema in effect says to strip off the predicate letter True and work with the result, taking care not to allow a "gap"-like formula to appear. If indeed A is grounded, the schema will determine it to be true or false [True("A") or True("$-A$")] in analogy with Kripke's construction. If on the other hand the schema when applied successively never eliminates all occurrences of True, then no conclusion of this form is reached, again in line with Kripke. This does not mean that negation of "True" is disallowed in our language. On the contrary; it is simply that the algorithm for calculating (equivalents of) True("A") proceeds by first eliminating negations of "True" inside A. That this has an intuitive meaning we see from Kripke: True("A") means A is grounded, first of all, and further whatever A says. In effect Gilmore's * tells us when it makes sense to "step back" from the meaning of a formula A in order to comment on that meaning.

Now, the above schema will be shown to be consistent in the following precise sense: If T is any first-order theory, then its first-order extension T' formed by including True as a new predicate letter, constants ("names") for all formulas A (which we usually write "A", although this may be ambiguous so that another mechanism such as Hollerith quotation is implicitly assumed), and the given schema supplying new non-logical axioms, is still consistent. Moreover, this can be proved by providing a model in a step-by-step fashion paralleling Kripke's construction. But now all the fancy metalanguage is left by itself in the usual Tarskian semantics, and the object language specifies all that we need to know about True. Thus True is defined in the very language to which it applies, and it is a total predicate in the usual first-order sense. As a consequence, we now will have True("A") or $-$True("A") for each A, although not True("A") or True("$-A$"). This means simply that the paradoxical cases *are* expressible in the theory itself, as paradoxes, yet without jeopardizing consistency.

Some peculiarity must be adjusted to for the Liar sentences L, where $L \leftrightarrow -$True("L"), since we will have neither True("L") nor True("$-L$"), and hence we will have $-$True("L") and $-$True($-$"L"). But L is a Liar, so from $-$True("L") we get L! Thus we must live with L and $-$True("L") together. But this is fine as long as we recall that now True is taken to mean Kripke's sense, i.e., grounded and true (in the Tarskian sense).

Now, what are we to make of a sentence such as L, for which we can prove both L and also $-$True("L")? Is L true or isn't it? Well, it certainly is *proven*, so in that sense it is established and in conformity with the facts of the situation; and it indeed will hold (be satisfied) in any model of the situation. It is only in the urge to call L "True" that we must restrict ourselves, and this is because of the very special nature of L, in that it itself specifically states that it is not to be so designated. Now this is the point: if we are going to allow our language sufficient flexibility to have variables that refer to expressions of the language itself, then such sentences as L will crop up. Our approach allows this, and recognizes them as the paradoxical sentences they are, without letting this create an inconsistency. The price is simply that we stick literally with what the sentences say, and this inconvenience will be as rare as are these sentences in typical discourse situations. We will give examples of this in Section 6.

5. The Consistency Proof

Let L be a first-order language. Consider an extension L' of L containing the predicate symbol, True, of one argument, as well as constants naming all expressions, such as provided by Hollerith quotation. It is natural to consider as an axiom schema for a theory over L' the following, for each term t naming a closed formula consisting of a string of symbols $e1 \ldots en$:

$$\text{True}(t) \leftrightarrow e1 \ldots en,$$

and which we ungrammatically write as True(t) $\leftrightarrow t$. (That is, t is a form of quotation, such as $n:e1\ldots en$ as seen earlier, and for simplicity we also write it in place of the string it names when outside the predicate True; alternatively, at times we will write the string in place of its name, thereby leaving off quotation devices.)

However, this can lead to Russell's heterological paradox, as follows: with only a modest amount of symbol-manipulation power, such as concatenation, one can construct a formula $R(x)$ that intuitively says that (the formula named by) x does not apply to its own name as argument, and then $R(R)$ will appear to assert its own denial in the form:

$$R(R) \leftrightarrow -\text{True}(R(R)) .$$

Definition 5.1. Call a *truth system* any first-order theory T having a designated constant $\langle e1\ldots en\rangle$ for every formula $e1\ldots en$, and a monadic predicate symbol True, such that for no term $\langle x\rangle$ are True($\langle x\rangle$) and True($\langle -x\rangle$) both theorems. (We usually will however omit the symbols $\langle$ and $\rangle$.)

Then by the construction of Russell's paradox we see that there does not exist a truth system (with concatenation) in which True(x) $\leftrightarrow x$ is a theorem for every closed formula x. (This can be regarded as a version of Tarski's "No Truth-Definition Theorem".)

As seen above, a key construction we will borrow from Gilmore is that of $x*$, the "positive" form of x: Call x positive if True is not in the scope of negation in the formula resulting from passing negation signs in x through to predicate letters, following the usual valid rules for this regarding quantifiers and connectives. (It is important for these purposes that conditionals $x \rightarrow y$ be written as $y \vee -x$.) Then let $x*$ be the result of passing "$-$" through True as well, so that $(-\text{True}(x))*$ is True($-x$). Then if x is positive, we have $x \leftrightarrow x*$.

Theorem 5.2. *If T is a consistent first-order theory, then T has an extension* GK(T)*—for Gilmore/Kripke—which is a truth system with axiom schema* True(x) $\leftrightarrow x*$ *for all closed formulas x.*

Proof (applying methods in [16] to ideas in [15]). Let $M0$ be a model of T, with domain D. Extend T by adding True and constant names to its language as above. $M0$ will still be a model of this extension if we interpret True as the null relation. We can regard $M0$ as determined by its true atomic formulas, i.e., those that hold there: these serve to interpret the predicate and function symbols. We will develop a model M of GK(T), where GK(T) is a truth system with axioms those of T plus the schema True(x) $\leftrightarrow x$ for positive x. We will do this by interpreting True in stages, starting (i.e., in $M0$) as the null relation, so that in $M0$ we have $-$True(x) for all x. As we extend the

applicability of True in further models Mu, we will be automatically determining new atomic formulas which are to hold. The idea here is that $\neg\text{True}(x)$ isn't necessarily permanent unless we first have decided $\text{True}(\neg x)$; the latter is regarded as definite once established, while the former may change as the sense of True grows.

Now, for any ordinal u for which Mu has been defined, let $Mu+1 = Mu +$ the set of "True(x)" for which $x*$ is true (holds) in Mu. That is, we change the interpretation of True in $Mu+1$ by making True hold for some additional strings.

This requires explanation. We suppose True to be part of the underlying language, so that True(x) does not hold in $M0$ as noted above. Then in $M1$, for each atomic truth in $M0$, such as $x = x$, we get True("$x = x$") as an atomic truth in $M1$ by definition, whereas in $M0$ we have $\neg$True("$x = x$").

For limit ordinals i, with Mu defined for $u < i$, let $Mi = \bigcup Mu$ $(u < i)$, where again Mu is regarded as represented by the set of its true atomic formulas.

Now the underlying language will have some cardinality k, i.e., the cardinality of its symbols, so also the number of formulas is k, and thus the sequence

$$M0 \subset M1 \subset M2 \subset \cdots \subset Mu \subset \cdots$$

must at some ordinal e become constant: $Mu = Me$ for all $u > e$.

Let $M = Me$. This is our candidate for a model of GK(T). (Note that GK(T) has no non-first-order rules of inference, so that this shall be a model in the usual sense.)

Since our goal is to show GK(T) is a truth system, we must show (True(x) & True($\neg x$)) is not a theorem of GK(T). This will follow if indeed M is a model of GK(T) and (True(x) & True($\neg x$)) is false in M for all x.

First, to show M is a model for GK(T), we need only verify the axiom schema True(x) $\leftrightarrow x*$ since the other axioms already hold by virtue of $M0$ (and hence M) being a model of T. So let True(x) hold in M for some x. Then True(x) already holds in some Mu, since it is atomic and $M = \bigcup Mu$ $(u < e)$. Assume u is the least such ordinal, hence not a limit. Then $x*$ holds in $Mu - 1$. But positive formulas, once true, remain so in our construction (this is a simple lemma) so that also $x*$ holds in M. Thus we have shown that True(x) $\rightarrow x*$ holds in M.

Now we turn to the converse: $x* \rightarrow$ True(x) in M. Let $x*$ hold in M. But $M = Me = Me + 1$, and True(x) holds in $Me + 1$ by construction. This gives the desired result.

We see then that True(x) $\leftrightarrow x*$ holds in M. Thus M is a model of the theory GK(T). Moreover, True(x) $\rightarrow x$ holds in M for all x, and we will need this fact to proceed. Observe that if we had $x* \rightarrow x$ for all x in M, then this result would follow. But in fact $x* \rightarrow x$ for all x in M. We see this as follows: Let x be

$-\text{True}(y)$, so $x*$ is $\text{True}(-y)$. Then if $\text{True}(-y)$ is true in M, then $(-y)*$ holds in M; but inductively assuming the desired conditional for fewer instances of connectives, quantifiers, and Trues than in x, we get $(-y)* \rightarrow -y$, hence also $-y$ holds in M. Now if $\text{True}(y)$ were true in M, then again we would have $y*$ in M and thus y as well, contradicting $-y$. So we have $-\text{True}(y)$, i.e., x, from the hypothesis $x*$. The more general case for arbitrary x follows similarly.

The above observation leads immediately to the conclusion that in M, $\text{True}(x) \rightarrow x$ for all x, since we have $x* \rightarrow x$ as well as $\text{True}(x) \rightarrow x*$. Now we see that $(\text{True}(x)$ & $\text{True}(-x))$ is false in M since otherwise we would have both x and $-x$ true in M. Therefore $\text{True}(x)$ & $\text{True}(-x)$ cannot be a theorem of $\text{GK}(T)$, and so $\text{GK}(T)$ is indeed a truth system. □

This indicates that Kripke's intuitive view of truth in terms of groundedness can be treated consistently in a first-order setting, so that excluded middle is upheld, and also the truth outcomes and the lacks thereof are all expressible within the formalism.

Since a Kripke-like model serves to show consistency of the Gilmorean scheme it follows that Kripke's intuitive sense of truth respects that scheme: that "true" sentences are ones that can be suitably tied to ground formulas. We have "improved" on Kripke, not so much in the meaning of the truth predicate as in the formal status (FOL) so that excluded middle is preserved and the "gaps" of Kripke become explicitly stated as $-\text{True}(x)$ (rather than simply failing to have $\text{True}(x)$ or $\text{True}(-x)$).

6. Sample Applications

A detailed treatment of applications to beliefs, concepts, and modal approaches to these questions will be pursued in a sequel. Here we consider some examples from ordinary reasoning, in which however belief and other cognitive notions are not at issue. For our first example, imagine Bill and Sue meet, and Bill begins the conversation:

"Did John talk to you about me?"

"Yes."

"Well, whatever it was, I'm sure it's a lie."

"But John told me that I can trust what you say."

This has various sorts of information being exchanged. We focus simply on the part dealing with truth, i.e., that Bill said that something, x, is false, and yet x itself is John's claim that Bill's statement is true. The fact that this is paradoxical is not what interests us at first; rather we are concerned with the representation of what is being said. In our formalism it is easily expressed:

Said(John,WJS) & Said(Bill,WBS)

& WBS = "Said(John,WJS) → False(WJS)"

& WJS = "Said(Bill,WBS) → True(WBS)"

where WJS and WBS are constants standing for "what John said" and "what Bill said", respectively. Since Bill does not know in advance what in fact John said, it is unreasonable to suppose John's statement to be represented as a concept at a given hierarchical level; indeed the example specifically illustrates the need for a lack of committment on this because as it turns out John's statement refers to Bill's, and so cannot support any ordinary hierarchy in which the outer statement (Bill's) must come either after or before the inner statement (John's): the two apparently must cohabitate one level.

Now to the paradoxical aspect of the example. A "naive" reading of True would in this context quickly lead to a contradiction. It is readily seen that True(WJS) implies True(WBS), and then False(WJS) follows. From this we conclude −True(WJS). Also we derive −True(WBS) for if True(WBS) then False(WJS) and so −True(WBS) after all. All this follows the Gilmore/Kripke interpretation. But now a naive reading would in addition infer True(−WJS) and True(−WBS) from the above, and the latter yields False(WBS) and so −False(WJS) from the definition of WBS. But True(−WJS) yields False(WJS), contradicting −False(WJS), and the paradox would thus deluge us in inconsistency and all its usual plethora of conclusions.

On the other hand, staying with the Gilmore/Kripke approach, we find only −True(WJS) and −True(WBS), harmless enough. Even if we replace False(WJS) and −True(WJS) in the example, we only derive, in addition to −True(WJS) and −True(WBS) as before, the conclusion WBS itself (unquoted). Then the paradox—WBS and −True(WBS)—is revealed but still harmlessly! WBS is seen to be of the Liar type, as indeed it is, but no contradiction ensues. The odd nature of WBS is not swept under the rug; it is represented faithfully and yet under control. Any effort to prove True(WBS)—which would lead to contradiction—simply goes around in a circle: in terms of the consistency proof, at no ordinal in the tower of models will WBS be decided to be true: −True(WBS) holds at every level because (WBS)* never holds. Indeed, we have WBS equivalent to −True(WBS). So WBS and −True(WBS) both hold at all levels.

Although it is not absolutely out of the question that a sufficiently astute use of hierarchies could deal with our example, it is by no means clear what this would be. We hope to have shown that a simple approach, without use of hierarchies or the abundance of separate languages that entails, is available. Straightforward efforts to represent the example so as to avoid self-reference (which is what hierarchies or separate languages would presumably avoid) seem fraught with difficulties. We illustrate this with further discussion of our example.

Suppose that Sue is a robot that has heard the utterances of Bill and John. If she represents them as we have done above, then of course our analysis remains unchanged. But suppose instead she regards statements by others as being in other languages than her own. Will this help her avoid a contradiction without the need for a special treatment of truth as we are urging? It seems not. For in order for her to understand what they are saying, she then must translate these utterances into her own language. Thus she may have the representations TrueB(WBS) $\leftrightarrow$ True(WBS′) and TrueJ(WJS) $\leftrightarrow$ True(WJS′), where TrueB and TrueJ are her predicate letters applying to "foreign" utterances in the languages of Bill and John, and WBS′ and WJS′ are her translations of these into her own language. Now, in order for her to reasonably be said to understand what she hears, she must also know the significance of their statements, i.e., that the truth conditions for WBS have to do with WJS, and vice versa, and she must be able to state this in her own language so as to reason about it:

$$\begin{aligned} \text{WBS}' &= \text{"SaidJ(John,WJS)} \rightarrow \text{FalseJ(WJS)"} \\ &= \text{"Said(John,WJS}') \rightarrow \text{False(WBS}')\text{"} \\ \text{WJS}' &= \text{"SaidB(Bill,WBS)} \rightarrow \text{TrueB(WBS)"} \\ &= \text{"Said(Bill,WBS}') \rightarrow \text{True(WBS}')\text{"} \end{aligned}$$

where equality here is equivalence modulo Sue's translation, and is surely something she must be able to utilize if she is to reason as we intend. But now Sue has reconstructed in her own language exactly the original form of the problem, and so will find the same issues as before. Unless she adopts a special treatment of truth she will derive a contradiction.

In case it may appear that we are deliberately and unnecessarily providing Sue with the ingredients for paradox, note that until she comes across that observation (of paradox) it is entirely reasonable for her to proceed as we have indicated. That is, until she has translated the utterances along the lines above, as far as she knows they may be harmless statements: Bill may have said John is a nice fellow, and John may have said that Bill speaks with a lisp; these statements can be regarded as having interrelated truth conditions also, but innocuous ones. Indeed, even their original statements are harmless separately, and it would be seem overly restrictive not to allow Sue to make sense of one person saying something about another's utterance. It is only after so doing that a conflict may turn up. It appears then that the ability to reason about (the truth of) what others say, carries with it the possibility of finding statements that refer to one another and hence are (at least indirectly) self-referential.

Now, a real person Sue in the robot's shoes would probably quickly go through some reasoning along the lines we suggest, and then smile at the paradox unconcernedly, reasoning that Bill mistakenly thinks John dislikes

him, and not bother further about truth assignments for the two original utterances: it simply won't matter to her, since she cares about other more significant information. Note though that it is not only a matter of focus of interest or attention. For until it is recognized that there may be a misunderstanding, either one of the statements might be true or false and it might be important to find out. Thus the determination, as well as the subsequent avoidance, of apparent contradiction is a necessary component of a formalization of such reasoning. The trick is to be able to observe the paradox and yet not be forced by it into an outright contradiction that could infect the whole reasoning process. Our approach does just that.

This is not to say that contradictions cannot be tolerated in a language for reasoning. After all, any number of contradictory pieces of information may be presented by various sources, and initially accepted. So other means of dealing with contradiction are also important. However, in some cases, as our example shows, it is useful to be able to discount the apparent contradiction immediately. Moreover, if the conflicting information is of the self-referential variety then it is not appropriately attributed to external errors of information; it is in the language itself and should be addressed as such.

In fact, at times it is essential to tease out the inter-referentiality of statements in order to draw a perfectly legitimate and desired conclusion. In the time-honored tradition of artificial intelligence, we turn to a logical puzzle for our final example. Consider the following problem of "Od and Id" offered in the spirit of Smullyan [17], although it is simpler than those he discusses.

> Forensic psychologist Jane Crane travelled to Lower Slobbovia where she was asked to solve a case of perjury involving two suspects named Od and Id. Now in Lower Slobbovia, humans always tell the truth. It was suspected that at least one of Od and Id was not human. Crane interviewed the suspects together, and the following statements were made:
>
> Id: "We both always tell only the truth."
>
> Od: "That's not true."
>
> From this information Crane was able to determine that Id indeed was not human.

We suggest that this can serve as a challenge for a formal reasoning system. It has an intelligible solution and involves no tricks. However, it does seem necessarily to involve statements that refer to one another to an extent that eludes a hierarchical approach. For consider how Crane might have proceeded:

> If Id is telling the truth then so is Od; but then Id's statement would not be true after all. So Id is lying (and so is not human).

The point here is that although the situation is totally improbable, it is one that people can reason about with success, deriving an unambiguous conclusion

by what seem to be ordinary modes of inference in ordinary language, unencumbered with caveats or contortions. We should expect a broad and flexible automated reasoning system to be able to do the same, or at least to be able to "follow" such a train of inference, and for this it must be able to represent the reasoning.

A representation of this reasoning in our approach could be as below:

$$\text{Human}(x) \;\&\; \text{Says}(x, y) \rightarrow \text{True}(y)\,,$$

$$\text{Says}(\text{Id}, \text{WIS})\,,$$

$$\text{Says}(\text{Od}, \text{"}-\text{True}(\text{WIS})\text{"})\,,$$

where WIS is an abbreviation for "what Id says", i.e.,

$$\text{WIS} = \text{"}(y)(\text{Says}(\text{Id}, y) \vee \text{Says}(\text{Od}, y) \rightarrow \text{True}(y))\text{"}\,.$$

It is now easy to formally carry out Crane's reasoning. From the hypothesis of True(WIS), WIS itself (unquoted) follows (given our treatment of True, for recall that True("A") $\leftrightarrow A*$, and $A* \rightarrow A$). Next from this and Says(Od, "−True(WIS)") follows True("−True(WIS)"), and therefore also −True(WIS). So true(WIS) → −True(WIS), hence −True(WIS) is proven. It now is easy to derive −Human(Id) from the first and second axioms.

We are not trying in this example to illustrate the power of the Gilmore * operator so much as the need to have a language in which mutually referring statements are expressible. The previous example also can serve, but it was paradoxical. Here we are arguing that also a straightforward and non-paradoxical conclusion can derive from such statements, and, we suggest, only when they are explicitly represented in a non-hierarchical fashion. Any attempt to untangle the mutual reference would seem to vitiate the information needed for Crane to reach her conclusion.

Note that in some sense a purely propositional account of the above is possible. For if we simply derive ourselves the needed components of the argument, namely, that if Id is human then WIS, and if WIS then −WIS, the conclusion is immediate. So the formalization

$$(\text{HI} \rightarrow \text{WIS}) \;\&\; (\text{WIS} \rightarrow -\text{WIS})$$

allows the conclusion −HI (not Human(Id)). Then of course there is no issue of self-reference. However, this misses the point, which is that it is possible to start from general information about people and what they may say, and then particularize it to cases. The propositional formalization just shown does not in fact contain the original information, and would not allow the further conclusion that, say Od is not human either if it becomes known that Od said Id is human.

7. Conclusions

The connection between self-reference and truth is simply that to do self-reference we need names for expressions, hence quotation (of some sort) and a way to relate the names to what is named, hence unquotation (i.e., a truth predicate). It is pointless, for example, to talk about beliefs outside the context of a world in which the beliefs may be true or false. So we become concerned with the relationship between Bel(x) and True(x). Obviously there is no hard and fast connection; that is not what is claimed at all. But there is a need to be able to represent the outcome (all cases arise: believed and true, believed and false, true and not believed, false and not believed). For this a language able to express its own syntax and semantics is necessary. We hope to have provided just this.

The unfortunate tendency to view first-order variables as rigidly tied to a narrow part of the world is probably due to the impoverished examples in logic textbooks, and leads directly to the outlook that second-order constructs are needed to talk about such things as the first-order variables themselves or their class of referrents. But in fact it is quite within the spirit of first-order logic to let (first-order) variables refer to syntactic and semantic features of the self-same language.

Indeed, the famous example of "all men are mortal" already implicitly expresses the correct attitude, in the formulation $(x)(\text{Man}(x) \rightarrow \text{Mortal}(x))$, where x presumably ranges over the whole universe, unspecified as that may be, except that it should include men. It needn't be restricted to organisms, physical entities, and so on; it can include societies, ideas, theories, and even all of these together. Moreover, we needn't say so in advance: we do not need to state that x above means all or any of some prepared list of "things". Rather, by means of axioms we make claims about certain values of x, such as the man-values; on the rest we remain uncommitted. Indeed the completeness theorem for first-order logic says that such a formula as above is derivable in whatever context of axioms we have, precisely if it holds in all interpretations (of the range of x and so on) in which the context at hand also holds. Thus unless stated otherwise by an axiom such as $(x)\text{Man}(x)$, the standard first-order semantics does not make restrictive assumptions on the "intended" range of variables.

Furthermore there is no reason, for instance as in LISP, that we cannot let a term stand for another term or even on occasion for itself. This simply is not seen very often, but is perfectly in the spirit of first-order logic and semantics. It does mean we will need a large supply of names for things, but this is no surprise; the hierarchical treatments also supply names (though in an extended language). This however involves issues of self-reference, and if given teeth with unquotation can then lead to paradox unless handled in an appropriate way such as we have indicated earlier.

Still, the consistency of the treatment of truth should not be taken as justifying the view that the beliefs of an intelligent reasoning system should be consistent. But our *theories* of its behavior should be consistent, and it may also fruitfully form its own consistent theories of its behavior, in which case it will need a way to refer to its own syntax and semantics. Our method provides just this.

Thus although Winograd [18] is right in that the semantics of a system can depend on properties of the processes involved, Hayes [19] is right in that (first-order) logic remains adequate to the task of expressing this dependency. As we discover more about the processes, we can express them in first-order logic, using quotation when necessary. The processes need not conform at all to the proof-theoretic mechanisms of logic, but can be whatever we deem appropriate; this has no effect on our expression of them as formulas of logic.

In conclusion, compunctions about free-swinging notation copied fairly directly from natural language, with its self-reference and relation-objectification (quotation), have kept us tied to overly weak and cumbersome representations ever since Bertrand Russell's discovery of paradox in Frege's theory of sets. Given our quotation mechanism, it is hard to see what serious restrictions are placed on knowledge representation by the requirement of first-order formalism. For we are more or less always restricted to discrete notations, and our efforts in natural language to express complex concepts rely invariably on object- and relation-terms. This means we can concentrate on the *facts* we wish to express regarding thought and action, and not be so concerned with novel mechanisms for expressing them. For what we wish to say about intelligence has straightforward expression; the difficulty is in discovering what those facts are.

ACKNOWLEDGMENT

This research has developed out of interests in foundational questions stimulated in me by two advisors, Martin Davis and James Allen, and further encouraged by a great many colleagues and mentors, including Chris Brown, Jerry Feldman, Alan Frisch, Andy Haas, Kurt Konolige, Henry Kyburg, Jack Minker, Nils Nilsson, and Dan Russell. It has been supported by grants from the National Science Foundation (MCS79-02971), the Alfred P. Sloan Foundation (78-4-15), and the University of Maryland (General Research Board Summer Award).

REFERENCES

1. McDermott, D. and Doyle, J., Non-monotonic logic I, *Artificial Intelligence* **13** (1980) 41–72.
2. Perlis, D., Language, computation, and reality, Ph.D. Thesis, University of Rochester, Rochester, NY, 1981.
3. Konolige, K., A first-order formalization of knowledge and action for a multiagent planning system, in: J.E. Hayes and D. Michie (Eds.), *Machine Intelligence* **10** (Halsted, New York, 1982) 41–72.
4. McCarthy, J., First order theories of individual concepts and propositions, in: J.E. Hayes, D. Michie and L.I. Mikulich (Eds.), *Machine Intelligence* **9** (Halsted, New York, 1979) 129–147.

5. Haas, A., Planning mental actions, Ph.D. Thesis, University of Rochester, Rochester, NY, 1982.
6. Elschlager, B. Consistency of theories of ideas, in: *Proceedings Sixth International Joint Conference on Artificial Intelligence*, Tokyo, Japan (1979) 241–243.
7. Weyhrauch, R., Prolegomena to a mechanized theory of formal reasoning, *Artificial Intelligence* **13** (1980) 133–170.
8. Attardi, G. and Simi, M., Consistency and completeness of OMEGA, a logic for knowledge representation, in: *Proceedings Seventh International Joint Conference on Artificial Intelligence*, Vancouver, BC (1981) 504–510.
9. Creary, L., Propositional attitudes: Fregean representation and simulative reasoning, in: *Proceedings Sixth International Joint Conference on Artificial Intelligence*, Tokyo, Japan (1973) 176–181.
10. Montague, R., Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability, *Acta Philos. Fenn.* **16** (1963) 153–167.
11. Burge, T., Epistemic paradox, *J. Philosophy* **81** (1984) 5–29.
12. Moore, R. and Hendrix, G., Computational models of beliefs and the semantics of belief-sentences, SRI Tech. Note 187, Menlo Park, CA, 1979.
13. Moore, R., Reasoning about knowledge and action, in: *Proceedings Fifth International Joint Conference on Artificial Intelligence*, Cambridge, MA (1977) 223–227.
14. Tarski, A., Der Warheitsbegriff in den formalisierten Sprachen, *Studia Philos.* **1** (1936) 261–405.
15. Kripke, S., Outline of a theory of truth, *J. Philosophy* **72** (1975) 690–716.
16. Gilmore, P., The consistency of partial set theory without extensionality, in: T. Jech (Ed.), *Axiomatic Set Theory* (Amer. Math. Soc., Providence, RI, 1974) 147–153.
17. Smullyan, R., *The Lady or the Tiger?* (Knopf, New York, 1982).
18. Winograd, T., Extended inference modes in reasoning by computer systems, *Artificial Intelligence* **13** (1980) 5–26.
19. Hayes, P., In defense of logic, in: *Proceedings Fifth International Joint Conference on Artificial Intelligence*, Cambridge, MA (1977) 559–565.

Received February 1982; revised version received September 1984

Languages with Self-Reference II: Knowledge, Belief, and Modality*

Donald Perlis
Computer Science Department and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, U.S.A.

Recommended by Patrick Hayes

ABSTRACT
Negative results of Montague and Thomason have diverted research in propositional attitudes away from syntactic ("first-order") approaches, encouraging modal formalisms instead, especially in representing epistemic notions. We show that modal logics are on no firmer ground than first-order ones when equally endowed with substitutive self-reference. Nonetheless, there may still be remedies, hinging in part upon a distinction between "dynamic" and "static" notions of provability and belief (an earlier version of this paper emphasized a somewhat different distinction).

Introduction

The focal point of this investigation is a result of Montague [23], whose customary interpretation as an argument in favor of modal logics for belief and knowledge as opposed to a classical first-order approach, I challenge. (For other responses to Montague, see [1, 4, 33].) I shall argue that modal logics are on no firmer ground than first-order logics when equally endowed with substitutive self-reference. Both modal and first-order treatments of knowledge and belief for commonsense reasoning can readily lead to inconsistencies. Yet there still may be remedies, depending on the particular forms of commonsense reasoning (and specifically of auto-epistemic reasoning) considered.[1]

Let us write Bel(x) and Know(x) to indicate that x is believed, respectively known, by an implicit agent g. The syntactic status of x is one of the issues to be addressed. If Bel and Know are predicate symbols, then x is an ordinary first-order term which in particular may be the name of a sentence.[2] On the

* This is a sequel to the paper "Languages with Self-Reference I: Foundations", *Artificial Intelligence* 25(3) (1985) 301–322.

[1] An earlier version of this paper [30] proposed remedies along somewhat different lines, that I now feel to be of less generality and usefulness.

[2] See [21] for an early call for naming, or reification, in AI.

Artificial Intelligence **34** (1988) 179–212
0004-3702/88/$3.50

other hand, if Bel and Know are modal operators, then x will be a well-formed formula. In [26, 28] it was suggested that for an intelligent reasoner g, a self-referential language is desirable in order to represent (to g itself) such notions as that g has a false belief. We may write, for instance,

$$(\exists x)(\mathrm{Bel}(x) \,\&\, \neg\mathrm{True}(x)) \,.$$

But if this very wff is a belief of g, then it too can serve (either in quoted first-order form, or in formula—modal—form) as an argument within another belief formula. I have contended [28] that this is such a basic aspect of language and thought that any reasonable representational mechanism for commonsense reasoning must include facilities for expression of self-reference and syntactic substitutions. We will see that this has significant consequences regarding consistency and modal treatments, in that apparent advantages of the latter over nonmodal ("syntactic") ones disappear in the presence of self-reference.

Now, the theorems of a proposed theory S for beliefs of an agent g can be viewed as themselves being the conclusions held by g, so that S is thought of as g's own reasoning context. Alternatively S can be viewed as the theory of someone, h, other than g, who is reasoning *about* g's conclusions. In the former case, g will be able to reason about g's own conclusions, and in the latter h may reason about g's reasoning about g's own conclusions. Of course, even more complex scenarios are possible, and have been considered in the literature, e.g., [4, 8, 14, 16, 19, 39]. In any case, such conclusions by g amount to beliefs of g. This will play a key role in our analysis.

A variety of theories has been considered for the study of belief and knowledge, many of them modal. S5, to be presented below, is perhaps the most famous of these. For now, I simply observe that S5 and other similar theories (modal *or* classical first-order) are very limited as theories of epistemic behavior of intelligent agents. In effect, they view knowledge (or belief) fixed once and for all in a timeless world; there are no processes, no mistakes, no guesses, no decisions, no plans, no goals, no new information. This is understandable: the idea here and elsewhere in formal studies has been to get a simple view right first, before turning to the complexities of real reasoning.

Still, the thrust of these theories has been one of deduction, that is, of additional beliefs an agent should come to, given certain original beliefs. There is in this an underlying, if not explicit, notion of process. While at times such complexities can be safely ignored for the sake of simplicity of analysis, at other times the very analysis can be impeded by such limitations. The more complex problem of real-time on-going reasoning actually suggests key ideas about the nature of knowledge and belief that cannot easily be seen from the more restricted idealized view. This will emerge in Sections 6 and 7. While we will not here endorse an explicitly process-oriented formalization of knowledge

and belief, the underlying idea of process will serve in our analyses. Key to this is the idea that an agent's set of beliefs may change over time.

Montague and Thomason showed that certain apparently plausible formalizations of the concepts of knowledge and belief have turned out to be internally inconsistent. I will recall some of these results here, and supply still more. Does this mean we must surrender formalism altogether, and look only at heuristic algorithms? I do not think so. Some of our results (Theorems 6.2 and 7.2) and an Open Problem point in directions that may be fruitful. What perhaps should be concluded is that timeless approaches may not be a good way to study knowledge and belief, or at least that timeless theories should be formulated *with an eye to* an underlying temporal framework.

My examination of formalizations of the propositional attitudes of belief and knowledge will lead to a distinction between two rather different kinds of theory I shall call "static" and "dynamic." To offer a brief preview, I call attention to the highly introspective feature present in theories of belief and knowledge. These are, largely, theories of *self*-belief and *self*-knowledge. That is, reasoning is specified so as to allow conclusions like Bel α or Know α from already having concluded α, and so on, in a succession of "layers." This layering however has been collapsed in most formalisms into a single flattened sea of conclusions. In the next section, I indicate in intuitive terms why this is suspect in general. Later sections give technical difficulties and potential solutions still within a flattened context. Roughly, dynamic theories, although flattened, are devised to take account of the (implicit) layers in a way congenial to commonsense reasoning, while static theories are not.

To set the stage for the rest of this paper, I briefly comment on the nature of knowledge and how it contrasts with that of belief. Indeed, the differences seem striking. Whereas knowledge is firmly tied to the notion of truth (as evidenced in the first-order schema Know('α')$\rightarrow \alpha$), that of belief is another matter entirely. In fact, knowledge has at times been characterized as simply true belief (or alternatively as justified true belief; but see [11]). In any case, there need be no presumption at all in agent h that a sentence believed by agent g is for that reason true. It may be true or false, and h's calling a sentence σ a belief of g distinctly raises the possibility of σ being false. That is not to say, however, that the agent g believing the sentence σ regards σ as anything but true.

Here I am taking the word "belief" in the (admittedly imprecise) sense of g's very definitely believing σ to be true, rather than merely suspecting σ to be true (as in "I believe so"). It may seem advisable to regard the beliefs of an agent g as simply some set or other of wffs. For some purposes this is too lax an approach; see [29] for a suggested narrowing of the definition of belief. However, as with knowledge, much of the literature has tended to treat belief as obeying very stringent "ideal" conditions such as logical omniscience (for example, obeying Bel$(\alpha \rightarrow \beta) \rightarrow$ (Bel $\alpha \rightarrow$ Bel β)). It is this approach that I

will explore here. (See [6–8, 17, 19, 26, 27] for exceptions.) We will find that many such approaches lead to inconsistencies, but eventually find others that are more promising.

The remainder of the paper addresses the following topics: Section 1 presents a sketch of the underlying intuitions motivating my analysis of introspective reasoning, which will crop up at a technical level again toward the end of the paper. Then Section 2 briefly reviews the importance of self-reference in commonsense reasoning, in particular with regard to formal substitution. In Section 3 general troubles with paradoxes of substitutive self-reference are reviewed, and modified consistent substitution-assertion rules are given based on [10, 28]. Then Section 4 reviews modal logics and problematic aspects of first-order analogues of modal logics for knowledge and belief in the presence of substitution à la Montague and Thomason, and in Section 5 we see that modal logics in the presence of substitution are problematic as well.[3] In Section 6 these problems are studied in terms of formal provability, providing a consistent static theory STAT of belief and knowledge; and in Section 7 a result of Löb as well as an example of Moore lead us to the aforementioned distinction between dynamic and static notions of provability and belief and suggest a dynamic theory DYNA. Then in Section 8, I summarize and suggest where more attention may be needed.

1. Flattened Layers of Introspection

Smith [34, p. 40] tells us "Perfect self-knowledge is obviously impossible . . . The self can never be viewed in its entirely, because there is no place to stand—no vantage point from which to look." In terms of our earlier remarks, the "obviously" here may mean that at best one can look back at ones "entirety" of a time prior to the present. There are layers of time to the phenomena of deduction and introspection. Flattening the layers into one is an enticing formal device, but a problematic one, whose only apparent usefulness is greater ease of study. But *certain* features of reasoning may persist invariant over layers and these may then be "flattenable," without the problems[4] that can beset this formal simplification. This is the issue addressed in this paper, and motivated in the rest of Section 1.

Example 1.1. If g's beliefs do satisfy some condition C, is it not reasonable for

[3] I caution the reader that this paper focuses principally on *first-order* modal logics rather than propositional ones, since we are interested in studying how self-reference is handled, and for this, quantification is needed, although the form that quantification takes is not necessarily that of first-order logic; see Theorem 5.3 and associated discussion.

[4] Compare [29] for a related use of problematic self-knowledge.

g to (come to) believe C itself? Yes and no. It may be that in coming to believe C, g's beliefs change in ways that may invalidate certain conditions, even C! As a simple example, if g has (initially) only two beliefs, A and B, then it is a fact, C, that g has only two beliefs. But if g then comes to believe this fact C about itself, it thereby ceases to have only two for it has now come by a third, C, which is therefore no longer true. On the other hand, if C were instead the fact of having *at least* two beliefs (rather than *only* two), then g's coming to believe C would not render C false.

Moreover, it seems clear that not all beliefs of an agent g *should* remain believed as new statements become believed. An obvious example is that the belief $\neg\text{Bel}\ \beta$ ("β is not one of my beliefs") should not remain if β comes to be believed. This is an example of what we can loosely regard as a *local perturbation* on g's belief set S. That is, many beliefs are contingent on small details, and can easily change.

Certain other beliefs, however, seem to be of more permanent character, such as

$$\text{Bel}(\alpha \rightarrow \beta) \rightarrow (\text{Bel}\ \alpha \rightarrow \text{Bel}\ \beta)\ ,$$

which asserts a global feature of g's reasoning processes. Of course, such a belief need not be true, or remain true over time. But it is perhaps plausible that *some* such general beliefs might remain true (for g) over a long period, and that g might come to believe *those*.

Herein lies the risk: will g's coming to believe a general belief γ about its very set of beliefs, alter that set in such a way that γ no longer holds, or so that it contradicts other beliefs of g? As with Example 1.1, it depends on γ. We will be concerned here to outline some broad categories of γ. But one thing we do want, is to be able to make local perturbations without thereby being forced to alter whatever global beliefs are held. Coming to believe β should not force giving up, say, $\text{Bel}(\alpha \rightarrow \beta) \rightarrow (\text{Bel}\ \alpha \rightarrow \text{Bel}\ \beta)$. Theories that tend to respect this requirement we call *dynamic*; ones that do not leave such global wffs invariant under local perturbations we call *static*.[5]

Another approach would be to openly embrace a hierarchy of theories, each looking back into the previous one. This latter suggestion we will not explore further here, although it is one that deserves attention (see [6, 9, 17, 29]). In this paper we will confine attention to single (flattened, ideal) theories of knowledge and belief, to see whether any can be made to avoid the kinds of difficulties mentioned.

[5] We will not be able to give necessary and sufficient conditions defining these terms. Nevertheless, they will serve our needs in directing our analysis and reformulation of various theories.

2. Self-Reference as a Principal of Language and Thought

In [26] the term *syntaxal* was used to describe languages having terms representing the syntactic features of that same language, the idea being that such self-descriptive power is essential for many purposes within natural language and commonsense reasoning. Rieger [32] uses the term *referenceability* for a similar notion, namely, that it is often important to reason (and communicate) about a particular feature of an utterance, viz., "That you said *howdy* struck me as unusual." Here elements of speech themselves are being referenced; for this some device is needed in a formalization. The quotation mechanisms presented in [10, 26, 28] are a possibility. An alternative device is found in modal logic, in which formulas (if not arbitrary expressions) are allowed to be operands to other expressions. One of our concerns here will be to contrast these approaches.[6]

We will use the expression "self-referential" as a gloss for either of these concepts (referenceability and syntaxality), noting that this leaves open whether all syntax is to be available for reference, as opposed, say, to whole formulas alone, or other aspects of language. This ambiguity will serve our needs, however, when we consider the alternatives of modal and first-order languages.

A further feature of natural language, and one that effective self-reference appears to hinge upon, is that of substitutivity. By this I mean the ability to refer to the result of making alterations in a statement, such as "If you had said *John is here* instead of *Mr. Smith is here*, I would have understood who you meant." The ability to form and compare such variations on our utterances is so elementary and fundamental to our use of language, that it is hard to imagine taking seriously any proposed formal language for a mechanical version of natural language processing that does not have a corresponding facility. Indeed, it seems tantamount to the ability to represent the very fact that language involves using symbols (e.g. *John*) to stand for other entities (e.g., John himself), as in "You used *Mr. Smith* to refer to John." Modal logic (as well as first-order logic) is in general broad enough to allow expression of such notions (if desired).

However, individual substitutions are not enough, as already suggested above in contrasting *John* and John. The very concept of substitution should be expressible, as in "If the subject is made plural, the verb should be also, so that if *people* is substituted for *person* as subject of a sentence σ, then the verb

[6] Boolos [2] and Smorynski [35] deal with another point of contact between self-reference and modal logic, namely, the use of a modal operator for the provability relation, and the corresponding treatment of self-reference as it derives from Gödel's Incompleteness Theorem [13]. This has connections with the present work, in that provability is one plausible model for a belief predicate as treated here; however, our focus is on surmounting certain paradoxes related to belief and knowledge rather than studying provability per se. See Section 6 below.

should be changed to its plural form." That is, the expression σ that is undergoing internal changes is not specified in detail, for a general rule is being given. This means that variables are needed to refer to expressions in the language. This amounts to little more than the ability to recognize symbolic strings, and so is not a computationally unreasonable condition to place on a language. Now here a modal logic in its usual form may need to be extended to allow variables for this purpose, whereas first-order logic does not require such modification.

Our starting point then is the contention that giving up such substitutions would be an unrealistic simplification of any formal language for commonsense reasoning. This would be analogous to a reasoning system g that behaves as follows: First, g makes assertion A and then is asked why it chose to make that particular assertion, instead, say, of a similar one with "John" used instead of "Mr. Smith." But g, having no concept of making a particular assertion made up of elements of some language, let alone of altering those elements, simply fails to comprehend what is asked of it. Of course, the design of full-fledged reasoning techniques to deal with such cases may involve many things; I contend however that an adequate treatment of self-referential substitution is one of them. Thus before turning to specific aspects of belief and knowledge, I explore some aspects of substitutive self-reference in first-order and modal logics.

3. Preliminary Results

We shall call a theory (over a language L) with mechanisms for expressing *and* asserting substitutions *unqualifiedly substitutive*. The hallmark of an unqualifiedly substitutive language is that it possseses an operator or predicate $\mathrm{Sub}(P, Q, a, n)$ directly asserting[7] the result of substituting in an expression P the expression Q for the nth occurrence of the subexpression a. I.e., if $P[Q/a, n]$ is the expression that results from the indicated substitution, then we are requiring $\mathrm{Sub}(P, Q, a, n)$ to be provably equivalent to $P[Q/a, n]$. Note that Sub here is to be an actual symbol (predicate or otherwise) of L, while $P[Q/a, n]$ is a meta-notation denoting some actual expression of L, namely the one resulting from the actual performance of the substitution. Of course, for the above-mentioned equivalence to be meaningful, the substitution must result in a well-formed formula of L.

It turns out that for the applications to be pursued here, a rather special variation on the Sub operator is required, namely one in which the substitution of Q for a in P be performed for precisely all occurrences of a in P except the

[7] That is, $\mathrm{Sub}(\cdot)$ is to behave intuitively as if it were $\mathrm{True}(\mathrm{sub}(\cdot))$ where sub is a function symbol. This will appear in more detail below. The reason I do not simply import True and sub wholesale here is that I have in mind various applications not all of which fit the mold of predicates and arguments (i.e., modal theories will have operators instead of predicates).

last. Therefore I will write simply Sub(P, Q, a). Contexts will vary slightly in that sometimes certain occurrences of terms will be quoted.[8]

As will be seen, the *asserting*[9] of the results of substitutions, i.e., relating the referenced syntactic elements to their intended meanings, runs into paradoxes of self-reference. Firstly, a means of unquoting quoted elements is needed, i.e., of saying formally that "α" carries the meaning α.[10] That is, Sub(P, Q, a) can be thought of as consisting of two conceptually distinct aspects: forming the new expression, and asserting it. These we can conveniently distinguish by writing the formula True(sub(P, Q, a)) where sub is a function producing (a name for) the expression that the indicated substitution leads to, and True asserts this expression. Again of course this can be meaningful only if the substitution leads to a wff of L.

However, this apparently cannot be done in such a direct way, for as Tarski [37] showed, the schema

$$\text{True}(\text{`}\alpha\text{'}) \leftrightarrow \alpha$$

leads, in any reasonably expressive language, to inconsistency.

The Sub concept described above is the key ingredient here. Unless care is taken, it will be possible to use the variables that range over expressions in such a way that they refer to that very symbol, Sub, and then it is often only a short step to paradox. I formulate this as a theorem in the first-order case; later I will present it as well for modal theories. In the present form it can be considered a variation on Russell's paradox as well as on Tarski's result above. Since Theorems 3.2 and 3.3 below amount to variants on ideas already present in the literature (in particular [10, 26, 28, 37]), their presentation here will be abbreviated, especially regarding quotation conventions. The reader can skim quickly through these preliminary results without loss.

For precision's sake I offer the following definition:

Definition 3.1. Let S be a first-order theory over a language L containing a three-place predicate symbol Sub together with the axiom schema Sub('P', 'Q', a) $\leftrightarrow$ P['Q'/a] where P['Q'/a] is as previously described, for all wffs P and Q and terms a of the language L (which is assumed to contain a constant 'α' for each wff α of L). Then S is said to be *unqualifiedly substitutive.*

[8] This is, again, since both modal and nonmodal contexts will arise. I have blurred details of quotes so as not to constantly write two forms for every predicate expression. The intention is always to have substitutions result in wffs of the appropriate sort for the context. I also have written $P[Q/a]$ for the result of substitution in either case.

[9] Here again I mean simply that we are using a formula with either an operator or predicate whose intuitive interpretation is to be that of the truth of its operand or argument.

[10] This is often represented as defining a truth predicate: True('α') is to tell us that the sentence "α" is true, so that True('α') and α should hold in the same models of a suitable theory.

Theorem 3.2. *Let S be an unqualifiedly substitutive first-order theory. Then S is inconsistent.*

Proof. We use $R(x)$ to abbreviate $\neg\mathrm{Sub}(x, x, y)$ and then $R(\text{`}R(y)\text{'})$ abbreviates

$$\neg\mathrm{Sub}(\text{`}\neg\mathrm{Sub}(y, y, y)\text{'}, \text{`}\neg\mathrm{Sub}(y, y, y)\text{'}, y) ,$$

which by the schema is equivalent to

$$\neg\neg\mathrm{Sub}(\text{`}\neg\mathrm{Sub}(y, y, y)\text{'}, \text{`}\neg\mathrm{Sub}(y, y, y)\text{'}, y) .$$

Here we are using the special substitution feature mentioned earlier, so that all except the last occurrence of y in $\neg\mathrm{Sub}(y, y, y)$ is replaced by $\text{`}\neg\mathrm{Sub}(y, y, y)\text{'}$. Thus we have $R(\text{`}R(y)\text{'})$ is equivalent to $\neg R(\text{`}R(y)\text{'})$, a contradiction.

An alternative (although less precise) argument is as follows: Define $\mathrm{True}(x)$ to be $\mathrm{Sub}(x, a, a)$ and apply the schema for Sub. This yields $\mathrm{True}(\text{`}\alpha\text{'}) \leftrightarrow \alpha$ for each α, which as mentioned above was shown in [37] to be inconsistent under fairly general conditions.

Part of what is required for the above kind of arguments is the mechanical ability to find and erase a symbol and put another in its place. As we saw above, this is fundamental to the expression of everyday (and significant) features of natural language and reasoning. Of course, all this depends on Sub having axioms that give it the intended meaning of actual symbolic substitution and assertion of the result, and so we can conclude that this is not possible without qualification, in a consistent first-order system.

In [10, 28] the difficulty of formalizing a truth predicate in first-order languages was circumvented, based on ideas in [12, 18]. Specifically, it was found that the above problematic schema can be replaced by

$$\mathrm{True}(\text{`}\alpha\text{'}) \leftrightarrow \alpha^*$$

for all sentences α, where α^* is essentially[11] the result of replacing in α all subformulas of the form $\neg\mathrm{True}(\text{`}\cdot\text{'})$ by $\mathrm{True}(\text{`}\neg\cdot\text{'})$. I will refer to this as GK (the Gilmore–Kripke schema).[12] GK is consistent in a broad setting. I am using the notation of [28] here.[13] It turns out that this approach can be applied fairly directly as well to the Sub predicate, and leads us to the following result:

[11] There are some qualifications regarding the form of α; see [10, 28] for details.

[12] This schema is adapted from Gilmore's work [12] on formalizing set theory, to capture ideas of Kripke [18] regarding truth predicates.

[13] However, I frequently abuse notation in that quotes and even parentheses may be left off. Thus $\mathrm{Bel}\ 0 = 1 .\!\rightarrow 0 = 1$ abbreviates $\mathrm{Bel}(\text{`}0 = 1\text{'}) \rightarrow 0 = 1$.

Theorem 3.3. *A ("qualifiedly substitutive") first-order theory S formed from extending a consistent*[14] *theory S' not involving the symbol* Sub, *by the addition of the (qualified) schema*

$$\mathrm{Sub}(`P', `Q', `a') \leftrightarrow (P[`Q'/a])^* ,$$

where now we define α^* *to be the result of replacing* $\neg\mathrm{Sub}(`P', \cdot)$ *by* $\mathrm{Sub}(`\neg P', \cdot)$ *in* α, *is consistent.*

Proof. First extend S' to S'' by adjoining consistently a function symbol sub and monadic predicate symbol True with axiom schema GK. Then define Sub as follows:

$$\mathrm{Sub}(x, y, z) \leftrightarrow \mathrm{True}(\mathrm{sub}(x, y, z)) .$$

It follows that this extension S'' is consistent, and clearly S is a subtheory of S''.

One thing I wish to investigate here (Section 5) is the extent to which the same result holds for modal theories. First I turn to a question addressed by Montague [23] concerning first-order analogues of certain modal theories.

4. Modal Theories and First-Order Analogues

The advantages of first-order logic over modal logic were pointed out by Montague [23] (I will review these later). However, Montague found that the "obvious" approach to using first-order logic instead of modal logic can lead to inconsistencies. Here I look again at Montague's results, to see whether the simplicity of his original suggestion for using first-order logic can be preserved somehow, and just why a modal syntax seems to manage what first-order syntax does not.

It will turn out that both answers are forthcoming, namely, we will see why modal syntax (sometimes) avoids contradiction, and we will be led to a better understanding of how to represent propositional attitudes in self-referential contexts (whether modal or first-order).

A modal language is characterized by modalities, i.e., operators that can be applied to formulas to produce new formulas, which however are not definable in terms of the standard propositional connectives. The most familiar examples are the necessity and possibility operators, sometimes written Nec and Poss, where usually one is defined in terms of the other, e.g., Poss α iff $\neg\mathrm{Nec}\neg\alpha$. Thus Nec α and Poss α are (modal) formulas, where α is any formula in the

[14] Actually we need a consistent theory with at least one infinite model, but this is a very minor restriction, for any consistent theory can be relatively interpreted in a consistent theory with infinite models.

language in question. Indeed, α may itself contain one or more instances of Nec or Poss or both. Thus a modal logic has an extended notion of formula, in which for any formula α and any modal operator M, $M\alpha$ is also a formula. In particular one can adjoin new operators (and axioms and rules) to a first-order language, creating thereby first-order modal logics. This is the case of primary interest for us.

It is a well-known result that the standard connectives (e.g., & and $\neg$) are sufficient to define all truth-functional operators in a propositional language, so operators that are not truth-functional are called modal to distinguish them from those already definable. Note that indeed the intuitive sense of Nec and Poss depends on more than the mere truth or falsity of their applicands. For instance, unless we are fatalists, a fire may have truly occurred in the house across the street yesterday, without it thereby being *necessary* that such a fire occurred there yesterday. On the other hand, we may feel inclined to say it was necessarily true that when the temperature of the house reached 451 degrees Fahrenheit, the books began to burn. In both cases, we may suppose the statement which is claimed to be necessary or not, is a true one. But in one case it merely *happens* to be true, and in the other it apparently *follows* from a general law about the ignition point of paper. This is not to say that this is a perfectly unambiguous distinction; nonetheless it serves to illustrate operators that cannot be treated as mere shorthands for expressions built of the propositional symbols. This has been taken as evidence that classical propositional logic therefore is inadequate to the task of representing non-truth-functional operators, and that modal logic should be introduced when such operators are needed.

In the hands of Hintikka [15] and Montague (see [25]), modal logics for representing concepts such as knowledge and belief have become powerful tools, and consequently a modal extension of first-order logic is regarded as a standard and natural representational medium for dealing with such matters. Thus once again the suggestion appears that extensions to standard logics are needed to represent appropriately the concepts of natural language, especially of belief and knowledge.

This unfortunately is not without its disadvantages. For one thing, first-order logic is much better understood than any modal logic formalisms, and consequently easier to apply coherently. Secondly, if some reins are not placed on the proliferation of new logics except when the latter are shown to be genuinely different (if not also useful!) from first-order logic, then we will end up with a tower of Babel, and research will probably suffer.

However, other avenues are open within first-order logic. One that appears promising is to use, instead of a formula as such, rather a quoted formula or term, so that the intended operator applies to such terms instead of formulas. That is, the operator becomes a predicate symbol: Nec('α') or Poss('α'). Then, so the hope goes, the corresponding axioms can be formulated satisfactorily without going beyond first-order logic.

This allows us to state a third technical benefit that would accrue from a first-order approach to propositional attitudes; in particular, in the words of Montague [23], "if modal terms [i.e., modal operators] become predicates, they will no longer give rise to non-extensional contexts, and the customary laws of predicate calculus may be employed." For instance, if in fact Bill and Kathy have the same phone number, a modal wff such as

Bel(John, phone(Bill) = 277-1265)

when coupled with Leibniz' Rule of Substitutivity (that equal terms may be substituted for one another without disturbing logical equivalence), yields

Bel(John, phone(Bill) = phone(Kathy))

even though John may *not* know this. Thus modal treatments combined with normal substitution practice is problematic, and special conventions are required to keep the unwanted consequences at bay. This suggests the attractiveness of remaining within a first-order language.

This is not to say that no problems remain in a first-order setting, of course. However, a first-order approach would instead involve the wff

Bel(John,"phone(Bill) = 277-1265")

which no longer has phone(Bill) as a term; rather the entire second argument to Bel is one constant term. Other similar difficulties that arise in substitution in modal contexts (sometimes referred to as opacity versus transparency of the modal operator in question), when treated instead via arguments that are quoted formulas, do not occur in first-order logic. The abandonment of first-order logic then is not to be taken lightly.

Motivated by these concerns, Montague [23] applied this approach to a modality for necessity. That is, writing Nec('α') instead of Nec α he obtained a quotational first-order construction. Montague proposed axioms for such a formulation, in analogy with standard axioms in the corresponding modal treatments. Unfortunately he found these versions to be inconsistent, whereas each corresponding modal operator version M is consistent. This seemed to be strong evidence in favor of the modal treatment. However, it appears that the inconsistency Montague uncovered hinges on certain fundamental expressive strengths of quotational first-order languages which are lacking in usual propositional modal languages. This is, first-order logics have richer sets of formulas than have traditional modal logics. Variables allow the formation of (self-referential) wffs that otherwise would not appear in the language, and thus more is being asserted in first-order logic than in the corresponding modal logic. The question then arises: if a modal theory M is made self-referential

(i.e., endowed with expression and assertion of substitutions), is it still consistent?[15]

One particular modal theory of interest is S5. Its language is that of propositional logic together with a modal operator. It was first studied as a formalization of the intuitive notion of necessity, with modal operator Nec[16], but also serves as a (tentative) formalization of the notion of knowledge. When I have knowledge in mind, I use Know instead of Nec, and when I have belief in mind I use Bel. In the immediate sequel I will employ Know. Note that α is a formula in a language containing Know, so that arbitrary nestings of Know are permitted.

S5 has the following axiom schemata:[17]

$$K\text{: } \mathrm{Know}(\alpha \rightarrow \beta) \rightarrow (\mathrm{Know}\ \alpha \rightarrow \mathrm{Know}\ \beta)\ ,$$
$$T\text{: } \mathrm{Know}\ \alpha\ . \rightarrow \alpha\ ,$$
$$I\text{: } \neg \mathrm{Know}\ \alpha \rightarrow \mathrm{Know} \neg \mathrm{Know}\ \alpha\ ,$$

as well as all (substitution instances of) tautologies, and the following rules of inference:

$$MP\text{: from } \alpha \text{ and } \alpha \rightarrow \beta \text{ infer } \beta\ ,$$
$$N\text{: from } \alpha \text{ infer } \mathrm{Know}\ \alpha\ .$$

S5 as presented above is a propositional theory, although it also has been

[15] It is of separate interest whether a first-order version of a modal logic can be kept suitably "weak" so as not to intrude, via its variables, new kinds of wffs that destroy a faithful match with the modal logic. This has been explored by des Rivieres and Levesque [33]. Our purposes here are somewhat different, namely, how to represent propositional attitudes in an explicitly self-referential context. Our contention is that apart from a desire to avoid inconsistency, there should be an underlying intuitive model justifying ones axioms. Then presumably whatever underlying intuitive model justifies the use of any particular modal formulation should apply as well to the full first-order formulation, unless that model itself indicates a principled argument to the contrary.

[16] Often written as L or a box $\Box$.

[17] In the literature, schemata K, T, and I are sometimes called Distribution, Knowledge, and Negative Introspection, respectively, and rule N is called Necessitation and sometimes written as RN. I is also sometimes called schema 5, since it is the distinguishing schema of S5. If I is dropped, the resulting system is called T (not to be confused with schema T, although schema T is the characteristic schema of the system or theory T). If schema T is dropped as well, the resulting theory is called K (with then characteristic schema K). Theory S4 has schemata K and T and the "Positive Introspection" or "PosInt" schema Know $\alpha \rightarrow$ KnowKnow α as well as rule N; so S4 is stronger than T; it turns out that PosInt is provable in S5 so that S5 is stronger than S4. Thus K, T, S4, and S5 are in increasing order of strength, and all have rule N and schema K; any such system (at least as strong as system K) is called (essentially) a *normal* system of modal logic. Another system, G, is studied in [2]; note that there "normal" is used as here except that rule N is not required. We refer the reader to [5] for more information on the (enormous) variety of modal systems studied. For a recent modal logic specifically designed for AI, see [20].

studied in a predicate context, with an underlying first-order language supplemented with the operator Know, and with the usual logical axioms and rules of inference as well as the axiom schemata K, T, and I, and rule N. It is in the first-order context that we are interested, so that in the remainder of this paper S5 will be taken to mean first-order S5.[18] Note that not only is the language extended beyond classical propositional or first-order languages, but also there is a nonclassical rule of inference. This requires comment. Rule N is to be applied strictly to actual theorems of S5, not to any hypothesis α we may wish to employ in proving, say, $\alpha \rightarrow \beta$. That is, although *if* α were a theorem of S5 then so would be Know α, this does *not* entail that the wff $\alpha \rightarrow$ Know α is a theorem of S5. In particular, extensions to S5 formed by adjoining new axioms are not assumed to obey rule N for any wffs other than the original ones provable in S5 itself, unless otherwise stated.

As a theory of knowledge, S5 has the following intuitive interpretation: Know α means α is known (by some thinking agent). Then K, T, and I may be plausible for an "ideal thinker" g. Schema K says that g's knowledge is closed under modus ponens; schema T that whatever g knows is true; and schema I that g knows whenever it doesn't know something (i.e., it can introspect negatively). Also, rule N is plausible if the agent is smart enough to know everything that can be established in S5, which may seem easy to grant for an ideal thinker. This presupposes that we are viewing S5 as *our* "external" theory *about* g's "internal" knowledge. However, rule N then has the further consequence that g will end up taking as its own knowledge all the axioms and theorems of S5 so that S5 ends up also being g's own theory after all. In this light, rule N serves as a kind of positive introspection mechanism.[19]

Montague studied several systems related to S5, with the particular aim of changing Know into a predicate symbol applied to names of formulas. I need not present details of these modal variants in order to state the following result of his:

Theorem (Montague [23, Theorem 3]). *Any first-order "arithmetical" theory having the schema* T', *namely*, Know('α') $\rightarrow \alpha$ *for each closed wff* α, *and also satisfying condition* N', *that* $\vdash$ Know('α') *whenever* $\vdash \alpha$, *is inconsistent.*

[18] See [2, 5] for more detail on S5 and its uses.

[19] The arguments to Know in S5 are interpreted as propositions rather than (quoted) sentences as in a syntactic first-order approach. This has some benefits, such as corresponding to Pierre's and Peter's knowing the "same" proposition that *Londres est jolie* and *London is pretty*, respectively. Also there are technically elegant semantics (due to Kripke; see [5]) that can be provided for S5 and other modal systems. However, this does not alter the fact that Pierre and Peter *express* their propositions sententially, nor that actual sentences are very important features of language. Thus our comments about the central role of self-referentiality and syntax seem to hold up. There are other qualifications regarding propositions as objects of knowledge or belief as well; for an overview of much of the philosophic literature, see [36].

Note that schemata *I* and *K* have been left out. I will henceforth however drop the prime on *T* and *N*, letting context determine whether a modal or first-order schema is meant. Also, I will retain the names *T* and *N* even when the predicate symbol is other than Know, e.g., Bel or the usual provability predicate Thm described later. The term "arithmetical" need not concern us; it is a gloss for a technical requirement that has the effect of allowing asserted substitutions into wffs.[20] We can establish an alternative theorem with a quick proof, if we stipulate a sub function directly, to get a variation on Montague's result that is more tailored to our needs. I begin with a definition.

Definition 4.1. If *S* is a first-order theory with function symbol sub of three arguments, and supplied with a distinct function term 'α' for each expression α (such that the free variables of 'α' are those of α) as well as axioms

$$\text{sub}(\text{`}P\text{'}, \text{`}Q\text{'}, \text{`}a\text{'}) = \text{`}P[\text{`}Q\text{'}/\text{`}a\text{'}]\text{'} ,$$

i.e., the name of the result of the indicated substitution, then *S* is *first-order self-referential*.

Theorem 4.2. *Let S be a first-order self-referential theory having a monadic predicate symbol* Know *and axioms* Know('α')$\rightarrow\alpha$ *for each closed wff* α, *and satisfying the condition* $\vdash$ Know('α') *whenever* $\vdash\alpha$. *Then S is inconsistent.*

Proof. Much as in the proof of Theorem 3.2, let $R(x)$ abbreviate the formula

$$\neg\text{Know}(\text{sub}(x, x, \text{`}y\text{'})) ,$$

so that $R(\text{`}R(\text{`}y\text{'})\text{'})$ abbreviates

$$\neg\text{Know}(\text{sub}(\text{`}\neg\text{Know}(\text{sub}(\text{`}y\text{'}, \text{`}y\text{'}, \text{`}y\text{'}))\text{'} , \text{`}\neg\text{Know}(\text{sub}(\text{`}y\text{'}, \text{`}y\text{'}, \text{`}y\text{'}))\text{'}, \text{`}y\text{'})) ,$$

which is equivalent to

$$\neg\text{Know}(\text{`}\neg\text{Know}(\text{sub}(\text{`}\neg\text{Know}(\text{sub}(\text{`}y\text{'}, \text{`}y\text{'}, \text{`}y\text{'}))\text{'} , \text{`}\neg\text{Know}(\text{sub}(\text{`}y\text{'}, \text{`}y\text{'}, \text{`}y\text{'}))\text{'}, \text{`}y\text{'}))\text{'}) .$$

So $R(\text{`}R(\text{`}y\text{'})\text{'})$ is equivalent to $\neg\text{Know}(\text{`}R(\text{`}R(\text{`}y\text{'})\text{'})\text{'}$. We further abbreviate $R(\text{`}R(\text{`}y\text{'})\text{'})$ as *RR*, so that we have *RR* iff $\neg\text{Know}(\text{`}RR\text{'})$. If we then use the

[20] Moreover, the use of substitution is virtually tantamount to the introduction of a certain amount of arithmetic in any case (see Quine [31]), and I have argued that substitution is an essential feature of commonsense reasoning.

axiom Know('RR')$\rightarrow RR$, we get

$$\text{Know}(\text{`}RR\text{'})\rightarrow \neg\text{Know}(\text{`}RR\text{'}),$$

so that Know('RR') is impossible and therefore $\neg$Know('RR') is proved. But this is (equivalent to) RR, so RR is proved. But then by the postulated inference condition, we deduce Know('RR') after all, a contradiction.

What does this result tell us? It appears that even a very weak subtheory of S5,[21] when "translated" into a first-order context, goes awry, at least in the presence of substitutivity. But is this reason to think that the modal version is better off? It is true that S5 (and therefore its subtheories) are consistent. But S5 by itself is not in a substitutive context. So the question arises as to whether modal theories such as S5 remain consistent when augmented with substitution capabilities.

In a similar vein, Thomason [38] has provided another apparent failure of our intuition, as follows: If an agent g believes (a suitable theory of) arithmetic and also g's beliefs (given as arguments to the predicate Bel) satisfy the following conditions:

$$\text{Bel}(\text{`}\alpha\text{'})\rightarrow \text{Bel}(\text{`Bel}(\text{`}\alpha\text{'})\text{'}),$$
$$\text{Bel}(\text{`Bel}(\text{`}\alpha\text{'})\rightarrow \alpha\text{'}),$$
$$\text{Bel}(\text{`}\alpha\text{'}) \text{ for all valid } \alpha,$$
$$\text{Bel}(\text{`}\alpha\rightarrow\beta\text{'})\rightarrow(\text{Bel}(\text{`}\alpha\text{'})\rightarrow\text{Bel}(\text{`}\beta\text{'})),$$

then g is inconsistent in the sense that g will believe all wffs.

To relate this to the intuitions of Section 1, I offer the following critique for Thomason's and Montague's theories as theories of omniscient ideal reasoning: In each case, an axiom schema (either T or the second schema above) attributes to g a global belief about g's own beliefs. This self-viewing is best taken as layered or steplike, and when instead it is flattened in one fell swoop then internal contradictions can arise. In effect, when g takes a position regarding the contours of its set S of beliefs, it may be embracing a "new" belief β which, if we force $\beta \in S$, can run into a self-referential dilemma. If g is aware of this, it might prudently choose to be more circumspect about its use of self-reference and in the process better merit our designation of it as "omniscient" or "ideal."

5. Substitutive Modal Logic

If we endow a modal logic M with the property of substitutivity in the form of an operator Sub(P, Q, a), with the intention that this thereby create suitable

[21] In fact all we need are schema T and rule N along with first-order logic and self-reference, so that not even the full system T is essential here, since schema K is not used.

conditions for referenceability within such an extended version of M, we have at least two available approaches. We can let P and Q be quoted expressions and Sub a predicate symbol, or we can let P and Q be formulas and Sub another modality.

Let us begin by exploring the first alternative. Since we already know that a first-order unqualifiedly substitutive theory is inconsistent (Theorem 3.2), so will be any modal theory M that extends such a first-order theory. Therefore, if we endow S5 with a predicate symbol Sub, we cannot allow it the unqualified substitution axioms as well. What then if we use only qualified substitution axioms of the sort known to be consistent in the first-order case? That is, can we extend S5 to include

$$\mathrm{Sub}(x, y, z) \leftrightarrow \mathrm{True}(\mathrm{sub}(x, y, z))$$

together with the consistent treatment of True and sub mentioned earlier, and thereby retain consistency in the modal theory that results? Unfortunately, the following result shows that we cannot.

Theorem 5.1. *If M consists of* S5[22] *extended by the* Sub *predicate with axiom*

$$\mathrm{Sub}(x, y, z) \leftrightarrow \mathrm{True}(\mathrm{sub}(x, y, z))$$

and associated qualified axioms for True *and* sub, *then M is inconsistent.*

Proof. We proceed by defining $R(x)$ to be the formula

$$\neg\mathrm{Know}\ \mathrm{Sub}(x, x, \text{`}y\text{'}) .$$

Then $R(\text{`}R(y)\text{'})$, which we will abbreviate as RR is

$$\neg\mathrm{Know}\ \mathrm{Sub}(\text{`}R(y)\text{'}, \text{`}R(y)\text{'}, \text{`}y\text{'}) ,$$

i.e.,

$$\neg\mathrm{Know}\ \mathrm{True}(\mathrm{sub}(\text{`}R(Y)\text{'}, \text{`}R(y)\text{'}, \text{`}y\text{'})) .$$

Now $\mathrm{sub}(\text{`}R(y)\text{'}, R(\text{`}y\text{'}), \text{`}y\text{'})) = \text{`}RR\text{'}$ and so

$$\mathrm{True}(\mathrm{sub}(\text{`}R(y)\text{'}, \text{`}R(y)\text{'}, \text{`}y\text{'})) \leftrightarrow \mathrm{True}(\text{`}RR\text{'}) ,$$

[22] An anonymous referee has pointed out that the proof does not use schema I, hence the theorem actually holds if S5 is replaced by any modal system with schemata T and K and rule N, i.e., for any logic as strong as modal system T.

whose right-hand side is equivalent to RR since True simply strips off quotes from its argument except when the symbol True itself is directly negated in this argument. What we have then is

$$\text{True}(\text{sub}(\text{`}R(y)\text{'}, \text{`}R(y)\text{'}, \text{`}y\text{'})) \leftrightarrow RR$$

and from rule N and schema K we then get

$$\text{Know}[\text{True}(\text{sub}(\text{`}R(y)\text{'}, \text{`}R(y)\text{'}, \text{`}y\text{'})) \leftrightarrow RR]$$

and then

$$[\text{Know True}(\text{sub}(\text{`}R(y)\text{'}, \text{`}R(y)\text{'}, \text{`}y\text{'}))] \leftrightarrow [\text{Know } RR] .$$

It follows that

$$[\neg\text{Know True}(\text{sub}(\text{`}R(y)\text{'}, \text{`}R(y)\text{'}, \text{`}y\text{'}))] \leftrightarrow [\neg\text{Know } RR] .$$

But RR is equivalent to

$$\neg\text{Know True}(\text{sub}(\text{`}R(y)\text{'}, \text{`}R(y)\text{'}, \text{`}y\text{'})) ,$$

and so we get that

$$RR \leftrightarrow \neg\text{Know } RR$$

Now, Know $RR \rightarrow RR$ by schema T, and so Know $RR \rightarrow \neg$Know RR, which means that $\neg$Know RR is a theorem of M. But we have just seen that RR is equivalent to $\neg$Know RR, so then RR also is a theorem of M, and by rule N so is Know RR, a contradiction.

We then consider the second alternative mentioned above, namely, that Sub(P, Q, a) be a modality in which P and Q are formulas. Here we are faced with a difficulty of syntax, if we wish to keep to our underlying premise in this paper, namely, that not only should language be substitutive and assertional, but that the very feature of substitutions should be expressible, in the form Sub(x, y, z) where x, y, and z are variables. This becomes problematic when x and y are intuitively to range over formulas (rather than names of formulas). What is called for is a quasi-second-order modal logic, in which the arguments to which modalities are applied (namely predicates, or more generally, relations) can be the values of variables. Thus we wish to write Sub(X, Y, z) where X and Y are predicate variables, and z is an individual variable that ranges over names of expressions. However, it turns out that it is not necessary to

adopt such a syntax, for even without variable arguments to modalities, contradiction arises.

Definition 5.2. S is an *unqualifiedly substitutive* modal logic if S has a modality $\mathrm{Sub}(P, Q, A)$ and the (by now familiar) substitution axioms using $P[Q/A]$, where P, Q and A are wffs. That is, $\mathrm{Sub}(P, Q, A)$ is equivalent to the result of substituting Q for all but the last occurrence of A in P. (We need not even use names at all, for instead of arbitrary expressions, it suffices to refer to whole formulas.)

Theorem 5.3. *Any unqualifiedly substitutive modal theory is inconsistent.*

Proof. We simply pick an arbitrary wff, say P, and consider the formula (1) below:

$$\mathrm{Sub}(\neg\mathrm{Sub}(P, P, P),\ \neg\mathrm{Sub}(P, P, P), P)\ . \tag{1}$$

The indicated substitution will then replace the first two occurrences of P in the first argument $\neg\mathrm{Sub}(P, P, P)$ of (1), with the second argument, which also is $\neg\mathrm{Sub}(P, P, P)$. This results in

$$\neg\mathrm{Sub}(\neg\mathrm{Sub}(P, P, P), \neg\mathrm{Sub}(P, P, P), P)\ , \tag{2}$$

which is in fact simply $\neg(1)$! That is, (1) is equivalent to $\neg(1)$, which is a contradiction.[23]

So S5 and even weaker systems such as T are inconsistent with either form of self-reference that naturally arises. Thus any advantage in a modal language seems to be lost, and we might as well remain with a classical first-order language. Still, this does not settle problems of formal representation of belief and knowledge. Whether formulated in terms of classical first-order substitutive or modal substitutive languages, special axioms and rules of inference for propositional attitudes are problematic. We now turn to remedies of this situation, hinging on separating the two mutually troublesome features, namely the schema T, $\mathrm{Know}(\text{`}\alpha\text{'}) \rightarrow \alpha$, and the rule $N$ for inferring $\mathrm{Know}(\text{`}\alpha\text{'})$ from α.

[23] It is of some interest that the requirement for "variable substitution" has been obviated by the very means of substitution. That is, in some sense the significance of variables is precisely that they allow for the possibility of substitution. This is not so apparent in first-order logic, where variables are central to the entire structure; but in modal logic where seeming arguments (to modalities) are not usually treated in argument fashion, the presence of substitutions is evidently just what brings things back to the first-order fold (at least in regard to self-referential paradox).

6. Belief as Provability

Our results indicate that modal logics, when endowed with sufficient power to represent substitutions, face the same inconsistencies found in first-order treatments. Much of the appeal of these logics is then lost, since one might as well then simply stay within first-order logic and employ a stratagem there for retaining consistency, instead of hunting for an analogous stratagem in modal logic. Indeed, we saw from Montague's and Thomason's theorems and from our Theorems 5.1 and 5.3, that there are severe difficulties whether the formal syntax is dressed in first-order or modal clothing. In effect, if we are going to have substitutivity of even a very mild sort we have to choose something less than full use of both the notion Know $\alpha. \rightarrow \alpha$ (schema T) on the one hand, and on the other hand $\vdash$ Know α whenever $\vdash \alpha$ (the familiar rule N).[24] What principled justification can be given for this, and what principled decision can be made toward a resolution?[25] One point of view emerges from the idea of provability.

Consider an agent g, whose conclusions are to be represented. Here I use the term "conclusions" as a deliberately neutral ground between knowledge and belief: whatever is in the agent's reasoning is to be used as if trustworthy. While this is still rather vague, it is sufficient for our purposes.[26] What we can say is that an agent's conclusions would seem to be tightly related to what the agent can prove (establish, decide, conclude), so that a natural idea is to explore ideas of provability in an effort to characterize possibilities and limitations on formal treatments of belief and knowledge. This idea was the basis for much of Konolige's efforts in [16], in which however provability was represented as a concept external to a given reasoner, i.e., one agent might reason about the provability relation of another agent but not about its own provability relation. In [17] Konolige comes closer to self-provability, but retains a kind of hierarchical approach in which what is provable at one level is then recorded as provable in the next. Here I am more concerned with a "flattened" theory that involves a predicate for its very own provability notion, yet in equal standing with the other predicates in the language; that predicate itself then forms part of the grist for those very proofs to which it is intended to refer. Boolos [2] and Smorynski [35] examine formal notions of provability, but in relation to arithmetic rather than epistemic concerns. In this and the next section, I study the extent to which various notions of provability are applic-

[24] I henceforth routinely drop quotes. All theories will be first-order (nonmodal).

[25] Asher and Kamp [1] pursue much the same question, and with a similar course by applying methods for truth predicates to a knowledge predicate. Their work involves a hierarchy of decisions about knowledge, and remains fairly close to Kripke's original idea [18] in that it is model-theoretically oriented. However, the needs of artificial intelligence (and possibly even of logic) seem better served by a more syntactic and proof-theoretic (i.e., computational) approach, as argued in [10, 28]. I will comment again on their approach below.

[26] I refer the reader to [29] for discussion of the issue of what constitutes a belief.

able to belief and knowledge. Initially I focus on the classical static theories; in Section 7, I turn to commonsense constraints.

Let us pursue the idea that an agent g's beliefs[27] are its theorems. But then if g is to have a Bel predicate of its own, it may be a kind of provability predicate. What do we mean, formally, that Bel be a provability predicate? Several things may suggest themselves. One, given in [3], follows.

Definition 6.1. A *provability predicate* for a theory S is a wff $P(x)$ that satisfies rule N—if α is a theorem of S, then so is $P(\ulcorner\alpha\urcorner)$—and also schemata K and S4's PosInt:

$$K: P(\alpha \rightarrow \beta) \rightarrow (P\alpha \rightarrow P\beta),$$

$$\text{PosInt}: P\alpha \rightarrow PP\alpha .$$

So, what is available for a provability predicate? There are many possibilities, most of which have little to do with provability. However, one provability predicate has played a special role in logic; it is due to Gödel, and we write it as Thm.[28] In this section I will focus on Thm. In the next section we will find that, for certain kinds of commonsense beliefs, this will not do, and we must examine alternative "introspection" predicates that are not provability predicates at all (as we have defined them).

Thm stands for the usual Gödel predicate symbol for provability in a "suitable" theory of arithmetic S, i.e., Thm is defined as

$$\text{Thm}(\ulcorner\alpha\urcorner) \leftrightarrow (\exists x)(\text{Proof}(x, \ulcorner\alpha\urcorner)),$$

where Proof(x, α) in turn formalizes in terms of arithmetic the proof-theory of S: it says x is (the Gödel number of) a proof of (the wff with Gödel number) $\ulcorner\alpha\urcorner$. Thm then pins down the mechanical details of what goes into a proof. This makes it static, for *no* new beliefs (axioms) can be adjoined now, without undoing the intended meaning of Thm. Thm lays out explicitly all the steps allowed in a proof, and even says that these are the only ones allowed. This explicit sense of Thm roughly corresponds to the specification of a mechanical listing of all and only the wffs that can be established by g (a recursively enumerable, but not recursive, set of wffs).

But how much of Thm is needed, anyway, in actual use of a belief predicate? A reasoner need not (and cannot) know *all* about itself, but might well benefit from knowing *certain* things about itself.

[27] We start here with belief, letting knowledge come in later. It turns out that knowledge is quite a tricky notion; see [11].

[28] Also often written as Prov or Bew (for Beweis).

In fact, Thm gives so much detail about proofs within a theory S that it inflexibly binds S away from any new knowledge. Thus if S is extended to S', the syntactic definition of Thm, if it is to now express proofs in S', must change. Moreover, certain facts will be new simply because they are not provable within S, and thus Thm cannot ever express them with reference to the theory for which it is formulated. In fact, the same happens with any provability predicate, as we shall see, regarding auto-epistemic knowledge of certain sorts.

Now, Montague's result shows that we must give up schema T, if we are to retain rule N and substitutivity (and consistency). That is, not *all* wffs of the form Bel α $.\rightarrow \alpha$ will be theorems of our agent g's theory S. What does this mean in terms of Thm? Intuitively, Thm $\alpha \rightarrow \alpha$ means that each theorem α of S is true (in a fixed standard model). Now, if S really has such a model, then each of these statements Thm α $.\rightarrow \alpha$ is correct; that these cannot all be *provable* in S is an interesting limitation of a computational mechanism to fully express its own computational behaviour. This is closely allied with Gödel's Second Incompleteness Theorem and also Löb's Theorem, discussed later. However, the underlying idea of the limitation has an intuitive sense to it, and will be the basis for the general picture that will emerge. The idea is that of Section 1, that the actual processes of a reasoner g can never be known in full detail by g itself, except by g's gaining this as new information which in turn changes g's structure so that what g has gained is a faithful picture of what it *was*, not what it *is*.

As I have stated, Know α is often taken to mean α is among those beliefs of g that are true. Then Know α means *to* g that α is one of its true beliefs, even though in general g cannot identify which these are! Indeed, each of g's beliefs is individually believed (by definition) by g; as soon as any one is seen to be false, it is no longer believed.[29] So g cannot isolate its true beliefs from the rest; it simply can refer to them in the abstract, just as it can refer to its entire belief set. In effect, g may believe that (the extension of) Know is a proper subset of (the extension of) Bel, but can give no examples of the relative complement (i.e., a false belief)!

Nevertheless, using Know as true belief, we can now employ schema T so that it applies to Know rather than to Bel. This manages to get around some of the difficulties we have seen. Specifically, the following result provides one way to endow g with its own knowledge and belief predicates and yet avoid inconsistency. This is a static approach, so that g will not be able to accommodate new beliefs and yet retain the intended meanings of Bel and Know.

Theorem 6.2. *Let S be any consistent qualifiedly substitutive first-order theory*,

[29] Try to imagine a reasoner g having simultaneous beliefs Bel α *and* $\neg$True α, as in "I believe 1337 is prime, but it is not!"

not containing the symbol Bel. *Then there is a consistent first-order theory* STAT(S), *which is an extension of* S *having predicate symbols* True, Bel, *and* Know, *and obeying rule N for* Bel, *with axiom*

$$\text{Know}\ \alpha \quad \leftrightarrow \quad \text{Bel}\ \alpha\ \&\ \text{True}\ \alpha\ ,$$

where True *satisfies schema* GK.[30]

Proof. Let True be as in GK, and then let Bel be Thm and let Know be as stated, both as extensions by definitions. Rule N is automatic for Bel (inherited from Thm).

How much of S5 does this result give us? We get rule N for Bel; schema T for True and Know; schema K for Bel, True, and Know. Theorem 6.2 does not give schema I for any of True, Know, or Bel. However, we do get S4's positive introspection schema PosInt for Bel, since Thm happens to obey Thm $\alpha \rightarrow$ ThmThm α.[31]

Example 6.3. (Belief biconditionals and cats). A key technique implicit in Montague's and Thomason's results, as well as in Tarski's and Gödel's, is the Fixed Point or Diagonalization Lemma (e.g., see [22]). This allows us to find, given a predicate $P(x)$, a wff α such that $\alpha \leftrightarrow \neg P(\ulcorner\alpha\urcorner)$ is provable in a suitable theory S. Substitutivity (qualified or not) is one criterion that makes S suitable (see [31]). The principal application for us is the following: Let Bel be a monadic predicate symbol of a theory S. We may then let BB be such that S has the theorem $BB \leftrightarrow \neg\text{Bel}\ BB$ (which I refer to as the belief biconditional).

Let C be (a formalization of) the sentence "The cat is on the mat," e.g., On(cat,mat). That is, C is a plain ordinary wff without self-reference, and without use of the predicate symbols Bel or Know or True or Thm. Its truth then should be determinate, even if unknown. Thus g should be safe in concluding not only $C \vee \neg C$ and Know $C \vee \neg$Know C, but also Know($C \vee \neg C$) and Know(Know $C \vee \neg$Know C). This follows from Theorem 6.2, and

[30] In [30] a GK version of rule N was used for Know: $\alpha^{**} \mid$ Know α, where ** is the Gilmore operator applied to Know rather than to True. This has the unfortunate consequence that straightforward (nonparadoxical) instances of theorems of g were not provably Known by g. For instance, even many *harmless* theorems such as $\beta \vee / \neg\beta$ did not have corresponding theorems Know[$\beta \vee \neg\beta$]. Thus we have dropped this approach and retained the Gilmore technique for the predicate True alone. Similarly, GK is not very satisfactory as a rule for Bel, for it severely undermines the introspection properties. For instance, if $\neg$Bel β is a theorem, so should be Bel$\neg$Bel β; but GK will not provide this. Also, as for Know, theorems of the form $\beta \vee \neg\beta$ should have counterpart theorems Bel[$\beta \vee \neg\beta$], but again GK will not provide this in general.

[31] It is of interest that Asher and Kamp [1] similarly arrive at a (semantical) framework for Bel, in which schema K and PosInt are preserved but schemata I and T are not. However, as mentioned earlier, their treatment has no corresponding proof theory for direct comparison to our work.

forms an interesting contrast with the wff BB. Let S satisfy Theorem 6.2. Then

$$\text{STAT}(S) \vdash \text{Know}(C \vee \neg C)\,,$$
$$\text{STAT}(S) \vdash \text{Know}(\text{Know}\, C \vee \neg\neg\text{Know}\, C)\,,$$
$$\text{STAT}(S) \nvdash \text{Know}(BB \vee \neg BB)\,,$$
$$\text{STAT}(S) \nvdash \text{Know}(\text{Know}\, BB \vee \neg\text{Know}\, BB)\,,$$
$$\text{STAT}(S) \vdash \text{Bel}(\alpha \vee \neg\alpha) \text{ for all wffs } \alpha\,.$$

Thus tautologies seem to behave well with respect to Bel, and *ordinary* wffs such as C also do so with respect to Know. But some explicitly self-referential wffs like BB fail to do so. In short, STAT(S) seems to represent a somewhat reasonable static theory of belief and knowledge. But it does not satisfactorily answer our doubts about schema T for Bel. For this, we need to look more closely at what use beliefs are put to, in order to assess what role schema T might or might not be reasonably expected to play.

7. Belief as Introspection

Theorem 6.2 above provides a version of Bel (and Know, etc.) that will suffice for certain purposes, such as nesting of belief sentences. As such it may be fine. But more might be desired. Let us see what this might be, by returning to the issue of schema T. Now, we know we may not have the full schema T (for Bel) in g's theory S, if S is to obey rule N (for Bel). And perhaps not all instances of schema T (for Bel) are even *plausible* for g to conclude. But perhaps *certain* instances are both plausible and possible. So first we should ask whether commonsense reasoning has need for these.

Another desideratum arises from rule N itself, for this rule does not necessary apply to new axioms that may be adjoined to S, even though the idea of rule N as an introspection facility should require this. We would like to have a formulation of Bel that is dynamic, in that as g gains new beliefs, the broad characteristics of Bel do not change, and thus that the rules for Bel should also not change simply because g has some additional beliefs. That is, there should be a core theory of belief that is invariant under mere accretion of (at least some[32]) information. If rule N is to be in this core, then it must be applicable to new (local) beliefs that arise.

We lead into this by observing that a different approach to introspection than the "static" one of Thm can be conceived. In particular, as Löb's Theorem will show, certain wffs when adjoined to a theory S result in the meaning of Thm becoming out of date, in the sense that Thm will express provability in the

[32] That is, what we loosely called "local perturbations" in Section 1.

original but not the extended theory. This is because Thm is so pinned down by its arithmetical definition to exact procedures of proof, that it cannot refer abstractly or generically to the general concept of proof. It is statically tied to one and only one set of conclusions. On the other hand, a reasoner g may not even be aware of its precise algorithm for reasoning, and yet refer to it generically in ways that do not depend on details (and therefore may remain consistent with a greater variety of extensions).[33] That is, perhaps a kind of referring to one's theorems or beliefs can be made, without explicitly stating in detail just how they arise. Possibly then Thm is then too fine-grained and unrealistic for commonsense reasoning, both conceptually (who could ever know all their mental processes?) and formally (Löb's Theorem below).

Where does introspection arise in commonsense reasoning? One prominent place is in nonmonotonic reasoning, in which account is taken of *not* having (believing, proving) a certain wff. For instance, one can believe à la Moore [24] that one always knows (or believes) that one has an elder brother if this is in fact the case, without necessarily knowing a way in which that conclusion or belief (that one had an elder brother) would be arrived at.[34] This is then *generic* or *abstract* information about ones set of conclusions. Just how such a notion might be approached and how it should differ from Gödel's explicit Thm will be taken up below. The point however is that an intelligent agent g may not need to know in any great detail just how its mental feats are accomplished; certainly human beings are in this situation. In fact, we will see formal requirements virtually forcing us into this position when we try to give auto-epistemic weight to Bel.

Now, can we think of a situation in which g ought to believe an instance of schema T, i.e., of Bel $\alpha .\rightarrow \alpha$? Well, there are trivial cases, the ones in which α is already a theorem. These of course are instances which STAT(S) will also produce. So how about nontrivial ones? Yes, following (or rather reversing) the example of Moore. We can suppose g to believe "if I believe I have an elder brother, then I have an elder brother" even if g does not believe "I have an elder brother." This would be a kind of infallibility belief for g, but a special one regarding brothers, and it seems perfectly plausible that an agent might have good grounds for such a belief as this.

Moore's (unreversed) example itself is also of interest; it has the form: $\alpha \rightarrow$ Bel α. Although this is not an instance of schema T, it is also not obvious that it is consistent with STAT(S) as long as Bel remains a provability

[33] This raises a number of issues, only some of which will be dealt with in the remainder of this paper. One I will not touch is that of hierarchical "cycles" of reasoning over time, as the agent realizes more and more about its (changing) inference algorithm(s). However, this would appear to be a key one for future work. See [9] for very interesting results on hierarchies of theories of arithmetic, and [6] for an approach to commonsense reasoning viewed as a steplike process.

[34] Hence, from the *not knowing* (of an elder brother), one concludes the *not having*.

predicate. Of course, here too there are trivial (or vacuous) cases, when $\neg\alpha$ is a theorem of S.

Definition 7.1. A wff α is *simple auto-epistemic over theory S* if α is of the form Bel $\beta \rightarrow \beta$, or $\beta \rightarrow$ Bel β, or $\neg$Bel β, where β is in the language of S. (S may or may not contain the symbol Bel.) I name the three types: MAE is the set of wffs of the Moorean form $\beta \rightarrow$ Bel β, RAE of the reverse form Bel $\beta \rightarrow \beta$, and NAE of the negative form $\neg$Bel β. Note that MAE wffs include instances of PosInt, and all RAE wffs are instances of schema T. I will sometimes abbreviate "simple auto-epistemic" as "simple AE."

This is not to say that all wffs of interest in auto-epistemic reasoning must be of one of the three given forms. Far from it. However, these three types seem to be the simplest ones and arguably the commonest, and also plenty of questions arise even for them.

Note that Thm does not satisfy schema T—this is essentially Gödel's Theorem on consistency proofs, and also can be seen in Montague's [23, Theorem 3]. That is, not all wffs Thm $\alpha . \rightarrow \alpha$ will be theorems of g. However, a theorem of Löb carries this much further, so that not a single *instance* of T will be provable except trivial ones for which α itself is provable. Call a wff Thm $\alpha . \rightarrow \alpha$ *trivial* if α is a theorem of S. We suppose here (as throughout the paper) that S has sufficient substitution properties, in this case the Fixed Point Lemma referred to earlier.

Löb's Theorem [2, 3, 22, 35]. *If P is a provability predicate for a consistent theory S, then no nontrivial instance of schema T (for P) is provable in S.*[35]

Corollary. *No NAE or nontrivial RAE wffs are provable in S if P is* Bel.

Thus if Bel is formalized (defined) as Thm, then *no nontrivial RAE* wff is a theorem of S. This is stronger than Montague's result that *some RAE* wff will fail to be provable (i.e., schema T in its full form clashes with rule N). Now we are faced with the fact that *each* instance of schema T (except trivial ones) clashes with the Thm interpretation of Bel.

What is going on here? After all, if a system (or reasoner) g does happen to have beliefs or theorems that are true (in some standard interpretation for

[35] This is a very striking result that seems counterintuitive at first. One consequence is that *no* wff of the form $\neg P\alpha$ can be a theorem of S if S is consistent. This also arises out of Gödel's Theorem on consistency proofs, which is easily proved from Löb's theorem and conversely. Gödel's result—his Second Incompleteness Theorem—is that $\neg\iota(\exists x)$[Thm x & Thm $\neg\iota x$] is not a theorem of S.

which Thm has the standard meaning), then all wffs Thm $\alpha .\rightarrow \alpha$ will also be true in that interpretation. But why then cannot g be made to prove this, since it is true? Well, on our suggested intuitions from Section 1, g can *come* to do so (by being given this extra information), but in the process g will change (as a formal system) and Thm will then characterize what g was, not what g is. Put differently, the result of Löb tells us g cannot know Thm to capture precisely *its* means of drawing conclusions; in fact it will not as soon as g thinks that it does! Or in AI terms, an ideal[36] g can never fully catch up declaratively with its own procedures for drawing conclusions.[37] For instance, the wff $\mathrm{Bel}\,0=1 .\rightarrow 0=1$ will not, by Löb, be provable in S, even though it will be consistent with any reasonable such S. Now if we extend S to

$$S' = S + \mathrm{Bel}\,0=1 .\rightarrow 0=1\,,$$

S' will be consistent but then Bel cannot be a provability predicate for S' (unless S' is inconsistent). Thus provability predicates are static: their properties do not remain invariant over additions of even very modest new introspective information.

Note that $\neg P\alpha \vee \neg P\neg\alpha$ is a kind of consistency statement for a provability predicate P. So $\neg\mathrm{Bel}\,\alpha \vee \neg\mathrm{Bel}\,\neg\alpha$ can also be regarded as a kind of self-consistency belief. Now, while this may in general be too strong for a realistic agent g, still certain cases of it seem unassailable. For instance, $\neg\mathrm{Bel}\,0=1$ should be concludable by g, on the basis that $\neg 0=1$ and that $\mathrm{Bel}\,\neg 0=1$. Yet Löb precludes this for provability predicates. Thus we separate two studies: static provability systems, and dynamic introspection systems that allow for the incorporation of new beliefs while retaining invariant general or generic information about beliefs as a whole. The latter requires giving up provability predicates as models for Bel; in their place we substitute what we shall call "introspection predicates."

It is true that Löb's Theorem applies to more than simply Thm; *any* provability predicate $P(x)$ gives us $\vdash \neg P\alpha$ for any α. However, the two postulates for a provability predicate in addition to rule N, namely PosInt and schema K, are not ones we should necessarily assume for Bel. That is, even though they may be true about g's reasoning, they are not facts g will necessarily know about itself. We then have the following result, toward a dynamic theory of belief.

Theorem 7.2. *Let S be any consistent qualifiedly substitutive first-order theory not including the predicate letter* Bel *in its language, and let AE be any set of*

[36] I.e., consistent, logically omniscient, and knowing sufficient arithmetic.

[37] A similar notion is exploited in [29] to characterize certain forms of default reasoning. Also Konolige [17] treats a similar theme from a different formal perspective.

Moorean or reverse auto-epistemic wffs over S. Then there is a consistent extension DYNA(S) *in which all AE wffs are provable.*

Proof. Let M be a model for S, and then form a model M' of $S + AE$ by interpreting Bel as truth in M; this will satisfy all auto-epistemic wffs of AE. For given a wff Bel $\alpha .\rightarrow \alpha$, if α is true in M, then α (and hence Bel $\alpha .\rightarrow \alpha$) already holds in M'; and if α is not true in M, then Bel α is false in M' and so again Bel $\alpha .\rightarrow \alpha$ holds. For a wff $\alpha \rightarrow$ Bel α, if α holds in M' then it also holds in M and so Bel α is true in M', making $\alpha \rightarrow$ Bel α true in M'; and if α does not hold in M' then $\alpha \rightarrow$ Bel α holds trivially. Then the theory of M' extends S and has all wffs in AE as theorems.

Corollary 7.3. DYNA(S) *above can be taken to obey* DYNA(S) $\vdash$ Bel α *whenever* $S \vdash \alpha$.

Proof. The same model M' in the proof of Theorem 7.2 is a model of Bel α for all theorems α of S, and so again the theory of M' serves.

It is necessary to restrict S and AE as above, namely S must not contain the symbol Bel, so that AE will not contain wffs in the language of S.[38] If α were allowed to be any wff in the *extended* language including Bel, then we could easily create Liar-type wffs and run into Tarski's theorem.[39] But we have already done what Thm (or any provability predicate) cannot do, in having even one nontrivial auto-epistemic belief present. Note that in modelling Bel as provability for S, we have not made Bel the same as Thm, for Bel is in the language of the extension $S + AE$, not of S. That is, Bel is not Thm for $S + AE$, and so Löb does not apply to the extension. But by the same token, Bel then is not an introspection predicate in the extension either.

Do we get similar results to STAT(S) for cats and BB here? Yes, from Corollary 7.3, letting S have the GK schema for True, and defining Know α be Bel α & True α as before, we have:

. DYNA(S) $\vdash$ Know($C \vee \neg C$),

DYNA(S) $\nvdash$ Know($BB \vee \neg BB$),

DYNA(S) $\vdash$ Bel($\alpha \vee \neg \alpha$) for all wffs α in the language of S.

Some natural cases have been missed, since we are working with a restricted

[38] And thus Corollary 7.3 does *not* provide the full rule N for Bel.

[39] Or Montague's theorem; e.g., from MAE wffs we would get rule N, and from RAE wffs we would get schema T, making the by now familiar deadly mix.

language in which there is no nesting of belief or knowledge. Thus we still have not achieved our goal of making g highly introspective as to its own beliefs. In particular, rule N is present only in very restricted form, applied to theorems of S but not to the extension in which the predicate Bel enters the language. Is anything else lacking? What might we want, for Bel to be an introspection predicate? Several things may suggest themselves. However, among the simplest is what I shall call the *double-N* rule, NN, which amounts to our familiar rule N from S5 and its converse N^{-1}, that is, α is a theorem of g iff Bel α is a theorem of g. Thus g can recognize (prove) it has α as a theorem (i.e., g can prove Bel α), precisely when it really does have α as theorem. This makes Bel in some sense "correct." Whether such a g is still realizable as a purely first-order theory is another matter. We are using rules of inference (N and its converse) outside of standard first-order logics, unless the logic in question obeys NN as a consequence of its axioms (note, for instance, that schema T in a theory makes N^{-1} is redundant). Thm does happen to obey NN, although due to Löb this will not help us here.

Definition 7.4. $P(x)$ is an *introspection predicate* for a theory S if it obeys rule NN:

$$S \vdash \text{Bel } \alpha \text{ iff } S \vdash \alpha \text{ for all wffs } \alpha .$$

Lemma 7.5. *If P and Q are introspection predicates for S, then for every term t,*

$$S \vdash P(t) \textit{ iff } S \vdash Q(t) .$$

Proof. Trivial.

It might appear from Lemma 7.5 that at most one predicate (up to equivalence) could satisfy NN, thus forcing it to coincide with Thm. For NN seems to characterize fully just what atoms of its associated predicate (e.g., Bel) can hold. After all, Bel α will be forced on g as a conclusion whenever (and only when) α itself is a conclusion. This might then seem to limit our formal choices for Bel very severely, indeed perhaps force us back to Thm and the loss of simple AE wffs. However, this is not necessarily the case, and leads to an open problem below. Roughly, Löb might not ruin our chances at an auto-epistemic formalization for Bel, because there might co-exist more than one introspection predicate for the same theory.[40]

Now we might ask whether introspection should go even further than than

[40] The beliefs of g will still be (the same as) the theorems of S; but it is not obvious that this necessitates, say, $S \vdash \text{Bel } \alpha .\!\rightarrow \alpha$ iff $S \vdash \text{Thm } \alpha .\!\rightarrow \alpha$, despite Lemma 7.5.

our new definition. In particular, the following may seem reasonable to consider:

Definition 7.6. A theory S with introspection predicate Bel is *fully introspective* if whenever a wff α in the language of S is *not* a theorem of S, then $\neg$Bel α *is* a theorem of S.

A fully introspective reasoner then would always be able to tell correctly for every wff whether it believed it or not: $S \vdash$ Bel α or $S \vdash \neg$Bel α for each α.

Now, such a reasoner, if consistent and knowing sufficient arithmetic, would have a non-recursively-enumerable set of theorems. Still, it may be of epistemological interest to know that in principle such reasoning could be envisioned. However, it is not to be.

Theorem 7.7.[41] *Every fully introspective qualifiedly substitutive first-order theory S is inconsistent.*

Proof. We form BB as usual, so that $BB \leftrightarrow \neg$Bel BB is a theorem of S. Now either BB is a theorem of S or it is not. If BB is provable, then (rule N, from Bel's being an introspection predicate) so is Bel BB. But also from the biconditional we get $\neg BB$, a contradiction. Suppose then that BB is not provable in S. Then by negative introspection we get $\neg$Bel BB, hence (from the biconditional again) BB, and finally (rule N) Bel BB, contradiction. (Note that we used only rule N, not full NN, so actually we have a stronger result than the stated one.)

We seem to be stuck then with (at best) the more modest introspection notion of rule NN for Bel. To recapitulate: static notions of introspection as in Thm are subject to Löb's Theorem while auto-epistemic versions of Bel ought not to be. We can do pretty well in a static (Thm- and NN-based) theory of belief, *if* we avoid consideration of AE wffs (Theorem 6.2). And conversely we can do pretty well in an AE theory of belief *if* we avoid NN (Theorem 7.2).

It is getting both together that remains problematic. We offer as a "*Belief Doctrine*" that Bel should satisfy both NN *and* a wide variety of cases of MAE, RAE, and NAE. It is then worth seeing whether some introspection predicate (other than Thm) might achieve this. In short, what kind of theories DYNA(S) obey rule NN? We know that if Bel is such a predicate, and if it obeys (within a theory S) schema K, then by Löb it cannot obey PosInt and so will *not* coincide with Thm. In our terms Bel would be dynamic and generic, failing to correspond in detail to the actual reasoning mechanisms of g. But we have

[41] This was inspired by a conversation with Richard Weyhrauch, Sardinia, October 1986. This result is not necessarily negative. Commonsense reasoners may well be inconsistent [29], and yet have interesting formal properties [6].

argued that this is appropriate for introspective reasoning. I leave the existence of such an *AE*- and *NN*-based (but not Thm-based) Bel as an open problem:

Open Problem. What subsets of *MAE*, *RAE*, and *NAE* are consistent with *NN*?

Regarding *AER* in particular, we know of course $\emptyset$ is fine (take Bel to be Thm), and that using *all* reverse wffs is never so (Montague). In fact, we cannot include Bel $BB \rightarrow BB$ for the same reason. But is there any nonempty subset of *RAE* that is consistent with *NN*? If so, then there can be more than one introspection predicate for one and the same theory: a provability predicate will not be the only choice, despite Lemma 7.5.

In terms of our notion of flattened layers, certain global statements such as PosInt and *K* will not allow mundane local statements such as *RAE* to be present. If we "force" them into an extension, we simply end up changing the theory so that we are, after all, looking back at the earlier "entirety" rather than the present (new) one. The Open Problem then is asking how much local perturbation (individual instances of simple *AE* wffs) we can get away with and yet preserve a useful amount of globality in the form of rule *NN*. Note that we did get a *weak* answer, in the Corollary 7.3, since rule *N* is obeyed there for wffs α that do not themselves contain the symbol Bel, and this already gives a number of cases useful in commonsense.

8. Conclusions

When a formal language is endowed with self-referential capabilities, especially in the presence of unqualifiedly substitutive mechanisms, difficulties of contradiction can easily arise. This holds for modal as well as (pure) first-order logics. However, the features of self-reference and substitutivity appear fundamental to any broad knowledge representation medium. Moreover, when remedies are taken, the modal treatments seems to offer no advantage over the first-order ones, and indeed the latter carry advantages of their own.

One can argue that although an agent *g* cannot *know* its beliefs to be true, still they *might* be true by good luck (or by the clever design of the agent's reasoning devices by a godlike artificial intelligencer), and all *g*'s inference rules might be sound as well. But then, if *g* is an ideal reasoner, wouldn't it be appropriate for *g* to believe this too? Wouldn't such an ideal *g* be able to believe Bel $\alpha .\rightarrow \alpha$ for all α? The odd answer (which we have seen in Montague's [23, Theorem 3]) is: not if *g*'s beliefs are to be consistent, which of course they must be if they are to be true. But this can be seen as an overly bold flattening of an essentially layered concept of Know. Of course, Thm is also a kind of flattening, but one in which no new information is to be brought in, by the very concept of Thm which pins down precisely what is allowed.

In fact, Löb shows us even more: that schema T can be allowed for *no* α except trivial cases, unless we give up provability as the measure of belief; then PosInt and other vestiges of pinned-down provability must be left aside. But since auto-epistemic reasoning depends on (certain instances of) schema T, agents will have to rely on dynamic (perturbation-tolerant) reasoning about their own beliefs. They cannot fully introspect; in particular, they will have to rely on pragmatic means to tell, for instance, that they do *not* have a certain belief.

I have been occupied here in showing that, after all, a flattened picture of commonsense may be available, a once-and-for-all set of wffs closed under certain procedures. This is what DYNA(S) really does, and it does so by leaving out wffs that might force the meaning of Bel to no longer be an introspection predicate. That is, the tradition of research that I have been exploring throughout this paper is in the mold of finding a fixed set of conclusions that g can believe, *but* allowing a reasonable amount of introspective self-reference at the same time.

Perhaps ironically, the static and flattened provability predicates, such as Thm, which obey PosInt and schema T, and which were found to thwart some attempts at a commonsense view, are the ones that would force a major *change* in any agent dealing with them, in a cycle of ever-expanding interpretations of its growing mechanisms of proof. Thus our dynamic version of Bel is not really one for a real-time agent at all. The conclusion, then, is that people have sought fixed formulations for belief using what are intrinsically nonfixed notions of self-description. For a single (and hence flattened) theory of belief to be viable, it must deal with predicates that are tolerant of self-description in a context of simple AE wffs; such theories are what I have called dynamic. The real trouble is that if Bel is made to look at itself closely enough, then it ends up describing a theory different from the one being investigated. This is fine if a cycle of ever-stronger theories is the focus of interest. However, single theories are vastly simpler to study, and so it is worth seeing how far this flattened approach can be carried.

The formalization of (fixed theories of) knowledge and belief still faces conceptual difficulties, especially in the case of agents whose beliefs are closed under logical consequence. In particular, it is unclear whether rule NN can be made consistent with reasonable commonsense instances of schema T. But it also appears that the study of agents with limited reasoning power, as has been initiated in [6, 7, 8, 19, 26, 27] is in great need of further study. Key to those approaches is the *absence* of the rule of inference "from α infer Know α" (with respect to a deductive engine which is sound and complete). Although some of these latter efforts utilize modal formulations, our work here strongly suggests that this is more a matter of taste than any real technical distinction, and thus that it may be preferable to stick with a common formal language to facilitate comparison in future work.

ACKNOWLEDGMENT

This research was supported in part by grants from the following institutions: (1) U.S. Army Research Office (DAAG29-85-K-0177) and (2) the Martin Marietta Corporation.

I would like to thank the following individuals for discussions that prompted me to carry out the elaboration of the ideas presented herein: Ray Reiter, Nils Nilsson, Jack Minker, Jennifer Drapkin, Michael Miller, Rosalie Hall, Brian Haugh, Kurt Konolige, Dana Nau, Jim Reggia, Maria Simi, Jim des Rivieres, Hector Levesque, Richard Weyhrauch, Bill Gasarch, Barry Richards, and Ian Pratt. I would also like to thank Pat Hayes for encouragement, and an anonymous referee for many helpful comments.

REFERENCES

1. Asher, N. and Kamp, H., The knower's paradox and representational theories of attitudes, in: *Proceedings Conference on Theoretical Aspects of Reasoning about Knowledge* (1986) 131–147.
2. Boolos, G., *The Unprovability of Consistency* (Cambridge University Press, Cambridge, U.K., 1979).
3. Boolos, G. and Jeffrey, R., *Computability and Logic* (Cambridge University Press, Cambridge, U.K., 2nd ed., 1980).
4. Burge, T., Epistemic paradox, *J. Philos.* **81** (1984) 5–29.
5. Chellas, B., *Modal Logic* (Cambridge University Press, Cambridge, U.K., 1980).
6. Drapkin, J. and Perlis, D., Step-logics: An alternative approach to limited reasoning, in: *Proceedings ECAI-86*, Brighton, U.K. (1986) 160–163.
7. Eberle, R., A logic of believing, knowing and inferring, *Synthese* **26** (1974) 356–382.
8. Fagin, R. and Halpern, J., Belief, awareness, and limited reasoning, *Artificial Intelligence* **34** (1988) 39–76.
9. Feferman, S., Transfinite recursive progressions of axiomatic theories, *J. Symbolic Logic* **27** (1962) 259–316.
10. Feferman, S., Toward useful type-free theories, I, *J. Symbolic Logic* **49** (1984) 75–111.
11. Gettier, E., Is justified true belief knowledge? *Analysis* **23** (1963) 121–123.
12. Gilmore, P., The consistency of partial set theory without extensionality, in: T. Jech (Ed.), *Axiomatic Set Theory* (Amer. Math. Soc., Providence, RI, 1974).
13. Gödel, K., Über formal unentscheidbare Satze der Principia Mathematica und verwandter Systeme I, *Monatsh. Math. Phys.* **38** (1931) 173–198.
14. Halpern, J. and Moses, Y., A guide to the modal logics of knowledge and belief: Preliminary draft, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 480–490.
15. Hintikka, J., *Knowledge and Belief* (Cornell University Press, Ithaca, NY, 1962).
16. Konolige, K., A first-order formalisation of knowledge and action for a multi-agent planning system, in: J.E. Hayes, D. Michie and L.I. Mikulich (Eds.), *Machine Intelligence* **10** (Wiley, New York, 1982) 503–508.
17. Konolige, K., A computational theory of belief introspection, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 502–508.
18. Kripke, S., Outline of a theory of truth, *J. Philos.* **72** (1975) 690–716.
19. Levesque, H., A logic of implicit and explicit belief, in: *Proceedings AAAI-84*, Austin, TX (1984) 198–202.
20. Levesque, H. Foundations of a functional approach to knowledge representation, *Artificial Intelligence* **23** (1984) 155–212.
21. McCarthy, J. First order theories of individual concepts and propositions, in: J.E. Hayes, D. Michie and L.I. Mikulich (Eds.), *Machine Intelligence* **9** (Wiley, New York, 1979) 129–147; also in: R. Brachman and H. Levesque (Eds.), *Readings in Knowledge Representation* (Morgan Kaufmann, Palo Alto, CA, 1982).

22. Mendelson, E., *Introduction to Mathematical Logic* (Wadsworth, Belmont, CA, 3rd ed., 1987).
23. Montague, R., Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability, *Acta Philos. Fenn.* **16** (1963) 153–167.
24. Moore, R., Semantical considerations on nonmonotonic logic, *Artificial Intelligence* **25** (1984) 75–94.
25. Partee, B. (Ed.), *Montague Grammars* (Academic Press, New York, 1976).
26. Perlis, D., Language, computation, and reality, Ph.D. Thesis, University of Rochester, Rochester, NY, 1981.
27. Perlis, D., Nonmonotonicity and real-time reasoning, *AAAI Workshop on Nonmonotonic Reasoning*, Mohonk, 1984.
28. Perlis, D., Languages with self-reference I: Foundations. *Artificial Intelligence* **25** (1985) 301–322.
29. Perlis, D., On the consistency of commonsense reasoning, *Comput. Intell.* **2** (1986) 180–190.
30. Perlis, D., Self-reference, knowledge, belief, and modality, in: *Proceedings AAAI-86*, Philadelphia, PA (1986) 416–420.
31. Quine, W., Concatenation as a basis for arithmetic, *J. Symbol. Logic* **11** (1946).
32. Rieger, C., Conceptual memory . . . , Ph.D. Thesis, Stanford University, Stanford, CA, 1974.
33. des Rivieres, J. and Levesque, H., The consistency of syntactical treatments of knowledge, in: *Proceedings Conference on Theoretical Aspects of Reasoning about Knowledge* (1986) 115–130.
34. Smith, B., Varieties of self-reference, in: *Proceedings Conference on Theoretical Aspects of Reasoning about Knowledge* (1986) 19–43.
35. Smorynski, C., *Self-Reference and Modal Logic* (Springer, New York, 1985).
36. Stich, S., *From Folk Psychology to Cognitive Science: The Case Against Belief* (MIT Press, Cambridge, MA, 1983).
37. Tarski, A., Der Wahrheitsbegriff in den formalisierten Sprachen, *Studia Philos.* **1** (1936) 261–405.
38. Thomason, R., A note on syntactical treatments of modality, *Synthese* **44** (1980) 391–395.
39. Vardi, M., A model-theoretic analysis of monotonic knowledge, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 509–512.

Received October 1985; revised version received March 1987

PART V

SELF-REFERENTIAL ARGUMENTATION

The Review of Metaphysics, Vol. IV, No. 3, March 1951.

COSMIC NECESSITIES

BY PAUL WEISS

I

Logical and Existential Necessity

A necessary truth is one which cannot be denied without absurdity. The absurdity can be one of two kinds — *logical* or *existential* — having corresponding negations in the form of logical and existential necessary truths.

A logical absurdity is self-contradictory. It results when a term and its negation are conjoined. It is intrinsically impossible; one need not look outside it to know that it is impossible everywhere. It is not because every domain rebuffs it that it is unable to appear in any, but because it cannot attain the status of appearing anywhere at all.

An existential absurdity is not intrinsically impossible. It expresses a conceivable state of affairs. It is not self-contradictory. It can appear; it can be entertained, believed, asserted. It is absurd because its existence or meaning denies what is required in order that it be or have a meaning. It is absurd because it denies the possibility of its own occurrence.

It would be self-contradictory to conjoin the nature of a domain with the negation of what is existentially necessary in that domain. Accordingly, we can transform every existential absurdity into a true logical absurdity by conjoining to it the characterization of the domain which it requires for its being or meaning. It is possible also to derive an existential absurdity from any logical absurdity, by using one term to define the character of subsequent discourse and then attempting to speak of the other term as inside that defined realm. Thus, "a square is a circle" is logically absurd, but "a square is the uncle of a circle" is only existentially absurd. The one is self-contradictory, the other conflicts with the unexpressed fact that mathematics deals with non-living beings and that these cannot have uncles. We can obtain a self-contradiction

by conjoining to the latter proposition this unexpressed truth about mathematics, so as to get some such logical absurdity as, "A circle is the uncle of a square, though neither is living and only the living have uncles." On the other hand, we can obtain an existential absurdity from the self-contradiction, "A square is a circle," by deliberately delimiting our discourse to nought but squares and then asserting, "this is a circle."

There are two kinds of existential absurdity — *definitory* and *essential.* A definitory existential absurdity is one which is absurd only because the assertion conflicts with some alterable decision. "This is a square" is an existential definitory absurdity when we restrict our discourse to circles, either because this is what we are now experiencing, or because we have arbitrarily refused to discourse of anything else. "Squares are uncles of circles" is also a definitory existential absurdity, since its absurdity depends on a tacit definition of mathematics as concerned with figures which are incapable of having the relation of uncle to one another. A change in decision would free both assertions of their absurdity. Even if we were now experiencing a circle, "this is a square" could be acknowledged to be necessarily true of a real world lying beyond the appearing circle; even if we accept the usual definitions in mathematics, "squares are uncles of circles" could still be acknowledged as a necessary truth in, for example, a fictitious world.

It follows that every contingent truth can be converted into a necessary, definitory, existential one by acknowledging the conditions for its occurrence as essential to its being or meaning. Thus, "Jones is walking," under the conditions of present observation, is not merely true but necessarily true so long as I take this observational evidence to define the truth of what I know. Contingent truths are thus definitory necessary truths held apart from that which provide them with their meaning or being; necessary definitory truths are contingent truths united with the conditions which make them possible.

Definitory truths are necessary only because one takes account of some limited domain with respect to which the discourse is to be in consonance. No one of such truths is itself

philosophically important; at most it reveals the nature of a universe of discourse we now happen to insist upon.

Essential truths are important in their own right. They are truths whose denials conflict with the possibility of any kind of discourse, experience, existence, knowledge or value, and thus with the possibility of themselves. They cannot be avoided, as definitory truths may be, by a mere shift in definition, assumption, belief or custom. To deny them is to deny the very conditions which the denial requires in order to be the denial that it is. For example, it is sometimes claimed that the senses always deceive. If the evidence for this is something learned by relying on the senses, the claim is, quite evidently, essentially absurd. We cannot reject the senses without rejecting all evidence, favorable or unfavorable, obtained through the agency of the senses. Since the assertion that the senses always deceive is offered in a sensed world in the form of sensed words to a sensing reader or auditor, "The senses convey some knowledge" is evidently a necessary, essential truth whose denial presupposes the opposite of what that denial says.

It was Kant who brought the attention of the world to the fact that some propositions could be denied only by rejecting the possibility of any experience. Unfortunately, he took experience in a rather narrow sense, supposing for example that it was rigidly determined, in the face of the obtrusive fact that it is chock full of contingencies, surprises, novelties and free activities. He somewhat arbitrarily supposed that there could be only twelve truths whose denials were essentially absurd, though the very definition of such truths compels no restriction to a dozen. And he overlooked in his first Critique what he had to admit in the second and third Critiques — that there were essential truths in ethics and esthetics whose import was in no way modified by the fact that instances of them might not be found in experience.

Experience ought to be acknowledged to be as rich, as fluid and as flexible as the empiricists have made out that it is. We need not subscribe to Kant's architectonic. We can acknowledge essential truths in other domains besides that of experience.

II

Formal, Necessary Truths

There are four kinds of essential truths — (a) those that are formal and prescriptive, (b) those that are formal and descriptive, (c) those that are material and prescriptive, and (d) those that are material and descriptive.

a. "Being is multiple," is an essential, formal, and prescriptive truth. All monists and most theologians would deny that it was essential, the former holding that there never can be more than one being, the latter holding that there was a time when there was but one being. Most logicians would deny that the proposition is formal, referring as it does to being, a reality outside the realm of formalisms. And all empiricists would insist that the proposition offers at best a probable description, prescribing nothing. Despite these formidable and reiterated contentions we must, I think, insist that the proposition is essential, formal and prescriptive and that the monists, theologians, logicians and empiricists are so far mistaken.

A denial of the proposition stands over against the proposition, over against its own component parts, over against the speaker and the world in which it occurs. It is thus one of many. We cannot then deny that being is plural without presupposing that it is. Being could be one only so far as the denial of this contention was impossible, and therefore only so far as we also were unable to maintain it. To the extent that it is possible to say that being is one, to that very extent we know it is not one. It might be argued of course that being might be one, but that we were not able to know or affirm the fact. But this contention cannot be maintained. A being not over against and in opposition to others would be completely indeterminate, an empty possibility for real beings. It would be identical, as Hegel saw, with Nothing.

The proposition is formal. The being to which it refers is being in any form, in any realm; it is what is common to every constant, a variable of the greatest possible range.

The proposition is prescriptive. It defines any discourse regarding a supposed single isolated being to be necessarily false. It lays down a condition to which any being anywhere must conform. It does not describe what is the case, but imposes a condition on any candidate for existence in any domain, actual or fictitious.

"Knowledge is other than being" is also essential, prescriptive, formal.

It is essential, for its denial is absurd; to assert it one must treat it as apart from one, as something to-be-known. To deny that knowledge is other than being is thus to commit the essential absurdity of affirming one thing and denying the very condition for making that affirmation. We must always distinguish between knowledge and what knowledge is about. The function of language is not to repeat but to report.

The given proposition is formal. It refers to any kind of knowledge and to any kind of being and relates them as other than one another. It might be argued, however, that the proposition was material, on two grounds. It might be contended that any variable whose scope is restricted by anything other than the structure of the proposition in which it occurs makes the proposition containing it a material one. "Knowledge" is such a restricted variable. Or it might be contended that any proposition containing terms which could not be defined as constructs out of the elements of logic is a material proposition. The first argument is too powerful; it would deny that a geometry, because it deals with points and lines, could really be formal. It would demand that logic deal only with variables that take values which could be anything whatsoever. In contrast, we ought to hold that a proposition is formal if it relates variables in the most abstract possible way, and this is what the given proposition, by means of "other than" does in fact. The second argument is too weak. It rests on the questionable belief that the terms of all formal disciplines are elaborations of the terms of logic. The thesis has never been made plausible to the majority of mathematicians and logicians. It is highly questionable whether there has ever been a single mathematical term which has been properly defined by means

of logical elements alone. And if one were to grant that the thesis has been successfully maintained in some cases, it is quite evident that it has been abandoned in others. The thesis has been presented in its boldest form in the *Principia Mathematica;* but that work cannot define a number as used by mathematicians except by bringing in the extra-formal idea of "typal ambiguity."

Kant's categorical imperative: Act only on that maxim whereby you can at the same time will that it should be a universal law, is also a formal, prescriptive and essential proposition, as can be seen most readily when it is rewritten as: Good men are those whose maxims assume the form of unrestricted universality. Since the proposition relates any good man to any universal maxim by a relation of implication, it is formal in nature. Since it tells us what must be acknowledged as the unavoidable elections of good men and what must be the nature of men whose maxims are not universalizable, it is prescriptive in nature. It is also essentially necessary. To deny it is to offer a truth and to that extent to act as a good man. And it is to offer the truth as that which others are to accept so far as they are good. The denial is thus an illustration of what it is purporting to deny. Kant's categorical imperative in ethics cannot therefore be rejected. But it does not suffice, since there are necessary, *material* truths, both prescriptive and descriptive, which must be acknowledged before an ethics is possible. Kant stated the indispensable conditions to which every ethics must conform — a tremendous advance which is not to be disowned because it is only a step on the way to the final goal of ethical knowledge.

"The beautiful is admirable" is on a par with the other three truths listed — it, too, is formal, prescriptive and essential. It is formal in so far as it treats the beautiful and the admirable as variables and relates them by a formal relation of implication. It is prescriptive, determinative of the attitudes one ought to assume to the beautiful, and of the characteristic to be predicated of that which cannot be admired. That it is essential follows from the fact that its denial is offered as a portrayal of what is so and has therefore the beauty characteristic of any

graspable truth. It is offered for approval to others and thus exhibits in itself the truth that something beautiful is admirable. The denial, in short, presupposes in itself what it purports to deny; it is therefore essentially absurd.

Such propositions as "being is multiple," "knowledge is other than being," "act in consonance with unrestricted rationality," "beauty is admirable," refer to all possible universes and not merely to ours. They do not therefore seem to be genuine philosophic assertions, but rather pre-conditions, general forms to which such assertions must conform. Yet, no one but a philosopher asserts such propositions. They ought to be termed "philosophical" but distinguished from those philosophical truths which pertain to our universe and perhaps to no other. Let us designate those philosophical truths which are essential to every universe, actual or possible, as "preconditional philosophical propositions," and let us use the term "philosophical" without qualification, to refer to those truths which are essential to this universe and may not be essential to any other. The latter will evidently be material, not formal in nature, formal truths serving only to provide the general framework of the material truths which the philosopher is attempting to utter. Accordingly, necessary, formal truths whether *descriptive* or *prescriptive* are preconditional in nature.

b. It is one of the intents of Husserl's philosophy to provide a method by which formal, descriptive, necessary truths could be obtained. I know of no such truth that Husserl's method has laid bare. Nor was Peirce, in his independent phenomenological inquiry, apparently more successful. Perhaps Whitehead's descriptions of the basic properties of extension in his *Process and Reality* come closest to fulfilling the requirement. But it is not evident that Whitehead was successful. The formality and descriptiveness of his account is evident, the necessity is not.

Phenomenology is a significant preconditional philosophical enterprise, but it is still a virgin domain. We could profit from careful convincing phenomenological inquiries in the fields of ontology, epistemology, ethics and esthetics to supplement the truths given above. The Husserlerian method of bracket-

ing, the Peircean dispassionate survey, the Whiteheadean mathematical dissection are not altogether adequate. We can have little hope of getting necessary truths which are purely descriptive and formal in nature. It does not seem that we can do much more in phenomenology than to assume arbitrarily that what we are describing is anything more than what happens for the moment and for us to be so and so.

III

Material, Necessary Truths

Philosophy discourses of this universe. Its truths therefore are material. Since to deny what it says is to deny the possibility of the denial itself, its truths must also be essential and necessary.

Of necessary, material truths some (c) are prescriptive, others (d) descriptive. It is these two types that are left for us to examine.

c. "No existent can be completely perfect" is a material truth, saying something of the substantial nature of things in this universe. It is a necessary truth, as true yesterday as it will be tomorrow, as true in Brazil as it is in Spain, true in the laboratory no less than it is on the street. It underscores the fact of conflict and of change, the limitation in the power and influence of each particular thing, for these are characteristic of what is imperfect. To deny that "no existent can be completely perfect" is to deny the necessary existence of a plurality of beings, each limiting and being limited by the rest. Such a denial rejects its own possibility, for it stands over against its own contradictory, its component parts, its speaker and the world in which it occurs and thus exists only as an imperfect entity, limiting and being limited by other imperfect beings. Only a single, undivided totality, exhaustive of all reality, could be absolutely perfect. Were there such a perfect being, it could not be known; for to know is to stand apart from what is known and thus to be less than all there is. The assertion that absolute perfection exists destroys its own ground.

Were the given proposition descriptive, it would isolate some characteristic which everything does and perhaps ought

to possess. Were it prescriptive it would specify a characteristic which all ought and may possess. Since the proposition lays down a condition to which all beings must submit if they are to be part of our universe of thought or fact, it must be prescriptive rather than descriptive in nature. Without having to survey and describe the whole of reality, we know that no being ever was or will be perfect in it. Of all conceivable candidates for existence, one kind we know therefore cannot exist — a God or Absolute, if these are thought to be exhaustive of all reality. We have here an ontological argument for the non-existence of Anselm's God or Hegel's Absolute. Of all possibilities these alone cannot exist. Or, to put it another way, the ontological argument is valid so far as it concludes from the possibility to the existence of a number of unspecified existent beings. Existence demands of possibilities that they be realized in groups. To exist is to exist over against, to be one of many.

There is a similar material, prescriptive, necessary truth in the theory of knowledge: "No actuality can be exhaustively known." The proposition is necessary, for if an actuality could be exhaustively known, there would be no difference between knowing and being, and knowledge would cease to be "knowledge of" and thus knowledge at all. The proposition is prescriptive, determinative of the nature of those things which can have the status of being actual. That the proposition is material as well, is evident if one contrasts it with its counterpart — the formal, prescriptive, necessary truth discussed above: "knowledge is other than being." The latter refers to an irreducible logical difference between any possible item of knowledge and any possible being; the former tells us a material fact about real beings and real knowledge.

A consequence of the foregoing is the truth "I am and always will be ignorant." This is an essential truth. For an empiricist, however, it can be nothing more than a definitory or a contingent one. He must allow that, on his principles, there is some degree of probability that he might be omniscient tomorrow. He illustrates in a new way how an excess of modesty is but a disguise for arrogance and dogmatism. One would be more modest by not being so modest — and one would

be more correct. No man can be omniscient. Knowledge involves the distinction and unification of elements to constitute a judgment; that judgment has the status of an item of knowledge only as standing over against a being in which the distinguished elements of the judgment are one and undivided. Knowledge of a table involves judging "this is a table" and referring and contrasting that judgment with the table as a "this" and "a-table", one and undivided. To know is to acknowledge a real being as the counterpart of a judged articulation of it; it is to confront the judgment of a being with the being which the judgment purports to be about.

Similar material, prescriptive, necessary propositions are to be found in ethics and esthetics. Both of these are normative sciences, enunciating prescriptive truths. Both of them assert prescriptive truths that must hold of all men everywhere. And since those truths refer to actual men engaged in acts which are actually right or wrong, monotonous or creative, and which terminate in what is good or bad, beautiful or ugly, they are truths which are material rather than formal in nature.

"It is absolutely wrong to reduce the value of anything" is prescriptive, dictating what acts ought to be and what ought not to be approved or performed. It is material, referring to men confronting real things in this universe. It is also necessarily true. To deny it is to assert that it might be right or at least indifferent if the values of some things were reduced. The denial makes sense only if it be understood to mean that the reduction of value does not involve the impoverishment of any being. This would be possible if value were an adventitious accident whose removal affects nothing. But this cannot be maintained.

To suppose that values were adventitious would be to suppose that they were irrelevant accompaniments of things. They would not be values *of*, but only values *for* them, and thus would be value-predicates, not values. Only if such value-predicates could somehow bring about relations or wholes which were in fact valuable, would there be values which could be destroyed. But those values would not be irrelevancies; they would be those relations or wholes as possessing some degree

of excellence. To reduce a value is to change a *being* so that it is less good, has less worth than it could have.

It is better for the good to exist than for it not to exist, to be concrete and determinate than for it to be a bare indeterminate possibility, for as existential the good is made effective and capable of being endlessly multiplied. A reduction in value is the conversion of some value as determinate and concrete into a value which is indeterminate and abstract. Were the destruction of things good, the more indeterminate things were, the better they would be. If such destruction were indifferent, there would be no value either in what was determinate or in what was indeterminate. There would be therefore no value anywhere.

"The indeterminate is not less good and may be better than the determinate" is thus a way of stating a negation of our initial proposition. But it has the status of such a negation only so far as it is determinate, determinately rejecting a supposed determinate falsehood. If what is said were correct, there would be no loss and there might be a gain in value were it less determinate and thus if what it said were not exact, or if it were a possible and not a real denial. The very reality of the denial shows that what it says could not possibly be true. It is only because it exists as a determinate entity that it can offer itself as a desirable alternative to the given proposition.

d. A descriptive truth discourses of things directly, isolating some feature which they in fact possess. It is formal or material depending on whether or not the characters to which it refers are abstract structures, possibilities, rather than substantial, specific beings. It is necessary, and essentially so, if everything, including its denial, in order to have being or meaning, presupposes its truth.

"Each being is unique" is one such truth. It is evidently not prescriptive, for it does not classify or evaluate but instead speaks directly of the nature that beings have. It is evidently a material truth, since it discourses of what actually exists in our universe. What is not immediately evident is that it is necessarily true.

The denial of the given proposition is: "Some beings are not unique." If the denial were warranted, some beings would

be duplicatable. It would be possible to have two of them which differed in no way whatsoever except in that they were two. But they could not be two except so far as they were different in some respect, say in location, in function, in appearance, etc. But then those locations, functions, appearances, etc., could themselves not be duplicatable, for otherwise they would not suffice to distinguish the two beings. It is possible to maintain that two beings could be merely numerically diverse only if one denied being to the points of space and time, to functions and appearances, and yet affirmed that these non-beings sufficed to distinguish two beings, which in every other sense were identical. One would thereby be maintaining that the numerical diversity of beings was an extraneous fact about them. But then the beings would not really *be* two.

"Some beings are not unique" is expressed by an individual in an individual way; it is a denial which occurs on an unduplicatable occasion. If it were true there would be some duplicatable being. Such a being standing over against an individual speaker on a unique occasion would, as in that context, be unduplicatable. One tiny little unique being off in the corner of the cosmos would make all other things absolutely unique, were they not — as they are — unique on their own account.

We are now perilously close to nominalism with its disastrous conclusion that there are no universals, no concepts, no propositions and therefore no theory of nominalism. But — and this brings us to another necessary, descriptive, material truth — though every characteristic and expression is individual because infected and sustained by an individual, one can, by convergent individual acts of abstraction, obtain a single result from those different individuals. What are termed 'repeated instances of universals' are individual cases from which, by individual acts, a common terminus in the form of a singular, indeterminate possibility, can be obtained.

"All knowledge is abstract" is a similar truth. It describes a character of knowledge, and is thus descriptive. Referring as it does to acts of human knowledge, it is material. That it is necessary is evident from the fact that it specifies the formal necessary truth "knowledge is other than being" by explicitly

characterizing knowledge as abstract and implicitly characterizing being as concrete. In brief, it states the particular way in which knowledge is other than being.

It is possible to deny that "all knowledge is abstract." The denial would be true in a realm where objects were abstract and knowledge was a vital possession of objects by beings more concrete. The denial is absurd because it is offered as an instance of knowledge in this universe, and thus as itself something abstract, communicable, something less substantial than the knowledge of which it discourses. All our discourse presupposes a world more concrete, and the denial therefore cannot exist except so far as it is false.

"Every being has value" is likewise material, descriptive and necessary. It is material, for it speaks of the beings in this world in non-formal terms. It is descriptive, referring to some trait they all possess. It is also necessary. He who rejects it as something false, or not proven, reveals in his attitude and effort his acknowledgment of the value of truth. Now whatever exists in a world where one relevant being is valuable has value of some kind, for it excludes and qualifies and may sometimes abet or conflict with the valuable being — and most important, will in its own way illustrate such value-traits as unity, existence, power and order. Because the denial has value of some kind, the beings of which it discourses must have value too, for they are the denial in a more determinate form. And whatever else exists with those objects will have value too, excluded and excluding, existing and unified just as they are. The things which the denial presupposes have a value which the denial absurdly denies they have.

IV

A Method

A good deal of what is here urged will appear to be obscure, even after the novelty of the statements has been worn away. Part of the reason for that obscurity is the fact that the truths on which we have fastened — and particularly the material, philosophically necessary truths — have a greater

reach than those in which we are daily interested and do not therefore have the translucence characteristic of what is at the focus of attention. The truths seem to be pulled out of nowhere. There is no indication as to just why these should have been chosen. There is no indication of how another might himself find these or similar truths. A method is needed for obtaining philosophic truths. It must begin with what is familiar or available to everyone and move with surety to the material necessities on which philosophy pins its hope and makes a claim to being sensible.

A philosophy can begin with any empirical truth whatsoever. There is none too trivial to serve as a starting point. It can begin with "This is yellow" or "My cat likes milk," with "I read fairy tales when a child," or "It rarely snows in Los Angeles." If we tried to start elsewhere we would be forced to begin with some specially favored common sense truth or go outside the range of common sense. In either case we would actually start with common truths but make explicit only the result of moving from them by means of a principle of selection or a method of going beyond them. A philosophy ought to begin where the philosopher actually begins, and thus with the truths he knows every day.

Different empirical truths stress different phases of existence or knowledge. The first of the above propositions is concerned with a fact of perception, the second provides an interpretation of a number of perceived responses which the cat makes to milk, the third is based on a report of memory and the fourth on descriptions and accounts of others. No one of them is exclusively a perception, an interpretation, a memory report or a description, but their difference in stress is so marked as to justify their classification as different types of assertion.

The first task of the philosopher is to identify out of the host of truths he knows those which stress that phase of existence or knowledge with which he is concerned, and turn these into necessary truths by relating them to their bases, to the reality of which they are the articulations. Every truth can be shown to be necessarily true, but only a few of them are such that their denial is tantamount to a denial of the possi-

bility of discourse, speech or life. "It is wrong wantonly to kill my friend" has this merit. It is not a truth which merely happens to be embraced by all men because they happen to have adopted some voidable base; it is a truth on which they must agree because the base which makes it necessary is inseparable from their individual beings.

Though all men have a direct acquaintance with a base which cannot be voided without losing sanity and coherence, and though it lies behind every voidable base as more general and permanent, it is hard to focus on. Unless one has the gift or good fortune to isolate truths which rest directly on the unavoidable base one must be content to look forward to them as the fruit of a long and exclusive devotion to philosophic study. Philosophers are usually gray by the time they are able to see and say clearly what it is that every one already vaguely knows.

It is not necessary for a philosopher to start with truths which refer directly to an inescapable base, though if he can start off with such truths he can proceed with greater despatch than otherwise. In their absence and as a check on them it is desirable to move from the stage of identification to the stage of generalization. Generalization is the process of taking an empirical truth and eliminating the different specific constants in it so as to free the whole from arbitrary limitation. "I opened the door to get here" loses nothing of its truth when generalized to "somebody did something to achieve something else," or more boldly, to "an effect follows on a cause." And we can generalize, "It is wrong wantonly to kill my friend" to "It is wrong to destroy wantonly," or more boldly, to "It is wrong to minimize value."

By generalizing, something of the obtrusiveness, concreteness and significance of the original is lost, but the truth is preserved in a simplified form. Because the generalization frees the truth from arbitrary limits it is one which has a more direct bearing on the inescapable base than the truths from which it was obtained.

A philosophical truth is a general truth which is inseparable from an unavoidable base. The higher up one goes in a

process of generalization the more surely one isolates truths which cannot be denied without absurdity.

A generalization is obtained by transcending the limitations expressed in a particular instance. It may be highly abstract — a form emptied of specific content. Every generalization, however, also has an exemplification. It expresses not merely the form which a particular case illustrates but the nature of a domain of existence or knowledge which specific things and bits of knowledge fill out in fact. "An effect follows on its cause" is a generalization exemplified as the temporal structure in which all actions occur; "it is wrong to destroy anything of value" is a generalization exemplified as the structure relating values and an absolute right.

The more superficial the truth which one accepts at the beginning and the more ultimate the field with which one is concerned, the more often must the process of identification, generalization and exemplification be repeated. If we accept, for example, nothing more than the fact a book is on the table, or that I am sitting at my desk, we are forced to repeat the process many times before we can reach such a basic truth as that all beings endeavor, in partly opposing ways, to become more and more complete.

Philosophy is the art of destroying the habit which keeps thought inside bounds having no greater warrant than convention and tradition. But no man can so purify himself that he does not in the end reveal the grip his time and his fellows exert. The boldest and freest of spirits are all creatures of their day. The philosopher must attempt to become completely free of the unnecessary bonds which hem man in, but he can hope only to become freer than he was. The degree of his freedom is measured by the extent to which he can repeat the threefold process and end with an exemplification which is broader and at least as concrete as the exemplifications with which he had been concerned before.

The purpose of the philosopher is to know the world in which he lives in a stable illuminating way. His beginning, end, and method reflect his sense of what is important; they reveal something of his sense of values. Philosophy is neces-

sary truth cutting the universe at its proper joints; but it is also, and to the very same degree, biography, exposing the character of him who is intent on revealing what lies about him. We can know the nature of philosophy's beginning, the method it must use and the kind of results at which it should arrive. But the movement of philosophy, the enterprise itself, requires a man who makes himself when and as he finds the world, and shows us himself when and as he tells us what he finds. Philosophy is the adventure of a free spirit in pursuit of cosmic necessities.

PAUL WEISS

Yale University.

On the Self-Reference of a Meaning-Theory

by ROBERT J. RICHMAN

HARVARD UNIVERSITY

A CRITERION in terms of which a division is drawn between meaningful and meaningless statements is, for purposes of the present paper, a meaning-criterion. I shall use the expression 'meaning-theory' as short for 'statement of a meaning-criterion.' Now it would seem, prima facie, that if a meaning-

[Continued on next page]

theory is such as to fall into the class of statements which on the theory itself are meaningless, then the statement of the proposed meaning-criterion—whether intended as an assertion or as a proposal—is nugatory. But some analysts maintain that a meaning-theory should enjoy a certain "metalinguistic immunity," a freedom from self-criticism which is based on the fact that the theory occurs on a different level of language from statements of whose meaningfulness it offers a criterion.

A recent writer upholds the claim to such immunity in terms of an analogy to a weighing machine which "we do not expect to weigh itself." At least one recent analysis has provided arguments against self-criticism of a meaning-theory. Since I find it easier to discuss the adequacy of arguments than of analogies (although it is easy enough to show points at which the analogy stated above fails), I shall concern myself in this paper with a consideration of the argument presented. I do not think that the claim to metalinguistic immunity of a meaning-theory should be granted. But, be it noted, I do not thereby assert that any particular meaning-theory is self-stultifying; such an assertion can be warranted only by showing that a statement of the criterion is a member of the class of statements which are meaningless in terms of the criterion.

The argument I shall consider is that of Professor Arthur Pap.[1] The meaning-theory he defends is the "verifiability theory of meaning," but this fact is not essential to our discussion. Replying to the charge[2] that the verifiability theory is self-stultifying, Pap writes that ". . . any argument against [the meaning-theory] from its self-inapplicability would be irrelevant."[3] For, he argues, "The critic takes for granted that the theory makes an assertion about a class of statements of which it itself is a member and is thus self-applicable. . . . But most proponents of the theory would hesitate to make this assumption since they know that once we admit statements that make assertions about themselves we entangle ourselves in paradoxes."[4]

Now philosophers—especially latter-day philosophers—frequently make statements about statements. It would seem to be a necessary condition of the acceptibility of such statements that the statement itself does not constitute a falsifying instance of what the statement asserts (about statements). Thus, if someone asserts that all statements are false, it does not seem irrelevant to remark that his statement if true is false; hence it must be false (since assuming it false does not imply that it is true). It would not seem unwarranted to "take for granted" that the statement is self-applicable, since unless the language is used in an unusual way, the assertion is included in the range of its applicability.

We might, indeed, in order to avoid the danger of paradoxes, refuse to admit self-referential statements into our language. This would eliminate,

inter alia, all statements about all statements. And this *would* render the criticism of a general meaning-criterion "irrelevant," if we may consider criticism of an unstatable position as irrelevant.

A standard sort of reply to such an argument runs in terms of "duplicates" of the statement in question at different levels of language. Adopting this line of argument, Pap writes, ". . . the verifiability theory [his proposed meaning-theory] must be understood either to refer only to the language of the first order, or else to duplicate itself for each linguistic order."[5] We may ignore for present purposes the obscurity in the notion of a linguistic order apart from some formalized system.

Pap thus offers two alternative solutions to the problem of having a meaning-theory which is not self-referential. On the first alternative, the meaning-theory refers only to statements of a specified linguistic level. But if we can formulate, for every statement of this level, an equivalent statement of a higher level, then the meaning-theory will be ineffective. And clearly we can formulate such statements in a very simple way, namely, by placing any given sentence of the specified linguistic level in quotation marks and suffixing to the resultant expression the words 'is true.' The sentence formed in this manner will occur at a higher linguistic level than that specified, and will thus be immune to criticism in terms of the meaning-criterion. Thus, it is easily seen that to be effective a meaning-theory must have a counterpart at every level of language above the first.

Let us assume that this is so; let us consider a proposed meaning-theory "to duplicate itself for each linguistic order." Then the meaning-theory of the second level serves as a criterion of meaningfulness for all sentences of the first level; and in terms of this theory we may perhaps reject certain sentences of the first level as meaningless. But then we must consider sentences of the second level, including, of course, the meaning-theory of second order, in terms of the meaning-theory of the third level. And if in terms of this latter criterion the meaning-theory of the second level appears to be false or meaningless, then we do not seem to be justified in rejecting as meaningless sentences which fail to conform to the meaning-theory of the second level. Thus, we must reject this demand for meta-linguistic immunity.

We may conclude by remarking that, as a matter of fact, it is by no means obvious (and, indeed, seems obviously false) that all self-referential statements lead to paradoxes. Insofar as we are not trying to construct a formal semantic system,[6] therefore, it seems unnecessary to resort to such a drastic measure as the elimination of all self-referential statements from our discourse. It seems sufficient for most philosophical purposes to eliminate piecemeal such statements as have paradoxical consequences. And thus we may feel justified in discussing the correctness of a sentence about

sentences—as, for example, a meaning-theory—without, in general, being forced into what is, from the point of view of ordinary usage, the never-never land of the meta-meta...language.

Received March 4, 1953

NOTES

[1] *Elements of Analytic Philosophy* (New York: Macmillan, 1949) pp. 341f.
[2] Leveled by A. C. Ewing in "Meaninglessness," *Mind*, 46:349 (1937).
[3] *Elements of Analytic Philosophy*, p. 342.
[4] *Ibid.*, p. 341.
[5] *Ibid.*, p. 342.
[6] In this case we may wish to lay down a general principle whereby we can assure the avoidance of certain known paradoxes. But even for this purpose it would be well worth while to examine the sort of sentences of natural languages which do lead to paradoxes, with an eye to finding general characteristics of these sentences which are not also characteristics (as is self-reference) of some logically unobjectionable sentences.

Argumentation and Inconsistency

by Henry W. Johnstone, Jr.

I wish to discuss a serious problem that is raised by my own position concerning philosophical arguments. My position is that all philosophical arguments are *ad hominem*. By this, I mean that a valid argument against a philosophical thesis must exhibit that thesis as inconsistent with its own assertion or defense, or with principles that must necessarily be accepted by anyone who maintains the thesis. I regard any valid argument *in favor of* a philosophical thesis as fundamentally reducible to arguments *against* theses held by the limited audience to which the argument is addressed, so that the polemical argument is, in my opinion, the basic type [1].

The problem that I now face concerns the inconsistency which, on my view, any philosophical critic must exhibit between his opponent's thesis and what that thesis presupposes, if his criticism is to be valid. What the critic must do is not only to exhibit explicit assertion X and presupposition Y, but also to call attention to the *inconsistency* between X and Y. The difficulty here is in showing why the defendant need *acknowledge* the inconsistency. If he is under no obligation to acknowledge it, then clearly the criticism could not have been valid. Indeed, many a philosophical argument has missed its mark for just this reason, and unless there are cases in which there is an obligation to acknowledge an inconsistency, there are no valid philosophical arguments at all. But it seems unlikely to me that every rational being would be under an obligation to acknowledge such an incon-

[1] For a fuller account of my view, see *Philosophy and Argument*, The Pennsylvania State University Press, 1959.

sistency, regardless of his philosophical orientation. My own orientation in philosophy leads me to this doubt, for two reasons. In the first place, the very notion of inconsistency is itself subject to philosophical interpretation. Later on in this paper, I shall discuss three possible philosophical views with regard to the conditions under which a set of statements is inconsistent. Thus whether a person is under an obligation to acknowledge an alleged inconsistency may depend upon whether it *is* an inconsistency in his terms. In the second place, if every rational being is under obligation to acknowledge an inconsistency, then the inconsistency must be subject to verification by every rational being—as, for example, "$68 \times 31 = 2{,}108$" is subject to verification. But cases of inconsistency subject to verification by every rational being can occur only in formal systems, in which the axioms and rules of inference are unambiguously stated and it is irrelevant to consider the meanings of the symbols. But clearly the assertion of a philosophical thesis or point of view is not like this, for the meanings of the symbols in terms of which the assertion is made are clearly essential to the enterprise of making the assertion.

Yet are we to conclude that there are *no* valid arguments outside formal logic? If so, argumentation is no more than an instrument of suggestion—albeit a more civilized and subtle one than hypnosis. There are types of argument for which this characterization is not at all improper. But philosophy is supposed to be an affair of reason. This can be maintained only if valid arguments can occur in philosophy. So we must try to make sense of the kind of inconsistency which such arguments would have to exhibit.

This essay is an attempt to illustrate and clarify the problem I have just been spelling out, and to take a tentative step toward its solution. First, I want to consider a concrete case of philosophical conflict, consisting of arguments and counter-arguments that could not be properly analyzed except in terms of the notion of inconsistency. Then I shall make some suggestions as to how this notion might be construed without treating it either as independent of philosophical orientation or as a purely formal notion.

I want to discuss the relationship between two contrasting philosophical interpretations of the laws of logic. The first of these, which I shall call functionalism, asserts that the significance of the laws of logic is exhausted by the role of those laws in mathematics and other areas where deductive proofs occur. If there are versions of deductive proof calling for alternatives to the usual laws, then alternative logics can exist side by side, none with more right than the others. For in order to establish the propriety of any kind of logic, it is sufficient to show that it has a use in deduction.

Functionalism in this sense would be opposed by what I shall call realism, according to which no law of logic is ultimately acceptable unless it is intrinsically connected with the structure of the world. Generally, realism lends its support to the so-called "classical" version of logic, although as I shall point out later, it is by no means necessary that it should. Classical logic is characterized by it inclusion of certain logical laws such as the laws of the excluded middle, and when realism espouses this version, it may show its interest in the structure of the world by expressing the law of the excluded middle, and other laws that it regards as fundamental, in ontological terms. It may assert, for example, that to be is, *inter alia*, either to have a given property or not to have it —what violates the law of the excluded middle cannot exist. According to realism, the alleged alternative logics lacking the law of the excluded middle are actually only incomplete versions of classical logic, which indeed is presupposed by the very act of interpreting these alleged alternatives. Realism, it will be seen, inclines toward monism, while functionalism is hospitable to a plurality of logics.

I have chosen functionalism and realism in logic as examples because their opposition seems to me to have a structure shared by many other pairs of opposing philosophical positions. According to one member of each pair, the meaning of any experience of a certain type reduces to its form. Thus, we have behaviorism on which behavior has no meaning over and above its form, ethical formalism, according to which the rightness of an act is uniquely determined by its form, and formalism in aesthetics, as well as the logical theory that

I have called "functionalism." The other member of each pair asserts that meaning resides in something over and above form; the meaning of behavior, for instance, must be sought in the conscious or unconscious mental activity of which the behavior is merely the expression. What I shall have to say about functionalism and realism in logic, then, would apply *mutatis mutandis* to many other examples of philosophical disagreement.

Let us consider the arguments for functionalism and for realism. The functionalist may point to various types of deduction in which, in his view, various logics are involved. The considerations leading to the development of modal logics are a case in point. The relation between the premises and conclusion of a valid argument is pointed to as a situation in which the classical treatment of conditional or hypothetical propositions falls short. For we do not say that an argument is valid merely because not all the premises are true or the conclusion is true. We say that it is valid only when it is *impossible* for all the premises to be true while the conclusion is false. There is, then, a logic of impossibility, possibility, and necessity, which contrasts with the classical logic of truth and falsity. To say that there "is" such a logic, however, is not to say that such a logic reflects the structure of what exists; it is only to say that there is a domain of deduction that requires it.

Modal logic is an alternative to classical logic in the sense that it involves a supplementation of the classical laws. Intuitionistic logic, on the other hand, represents a weakening of classical logic. Yet the functionalistic argument can be used to defend intuitionism as easily as it can be used on behalf of modal logic; for there "is" a logic without the laws of double negation or excluded middle only to the extent that there are areas of mathematics demanding this logic. No doubt, a non-functionalistic defense of logics that deny the laws of double negation or excluded middle, is also possible. The Marxist defense is a case in point; in a world of dialectical flux, the negation of the negation is an emergent. This is actually what I have called realism; for it sees logic as the mirror of reality. Modal logic can undoubtedly also be defend-

ed on the basis of cosmic necessities and possibilities. For that matter, classical logic has a functionalistic defense; witness the formalism of Hilbert. What these variants show is just that one should not confuse types of logic with interpretations of logic. It is with interpretations that I am now concerned; for it is interpretations of logic, not types of logic, that can collide philosophically.

I have spoken of the functionalistic arguments for various types of logic. The evidence on which such arguments rest is simply the existence of various domains of deduction, each alleged to require the type of logic in question. I turn now, to the realistic argument for a logic based on ontology. The evidence here is the world. But it might be objected, for example, that the world does not constitute clearout evidence of the law of double negation. What is *solid* is *not-liquid*, but what is *not-liquid*, is not necessarily *solid*. So the denial of the denial of *solid* is not necessarily *solid*. Supposing that the realist wishes to defend classical logic, he will reply that this objection rests upon a misunderstanding of what it means to deny a term. Once this point is cleared up, the formula "For any property P, P = non-non P" expresses without distortion an aspect of the structure of the world.

But there is something unsatisfactory about both the functionalistic and the realistic arguments that I have outlined. They seem to sidestep issues that one thinks ought to be met head-on. They certainly do not meet each other head-on. The outlines I have given are, in fact, incomplete; and what is missing will soon be obvious enough. As matters now stand, both functionalism and realism constitute excellent documentation for the statement that every philosopher has reason on his side and none can, accordingly be refuted. The elaboration and defense of each doctrine is clearly a rational activity. For the functionalist, it is the activity of examining various domains in which deductions occur to see what logics are required. For the realist, it is the activity of correcting language, so that the ontologically warranted logic can appear without distortion. But no functionalist will be much impressed by the report that there is ontological evidence for the ultimacy of a certain type of logic; the "evidence" shown him

is simply not evidence for him at all. Nor does his own report that, for example, certain areas of discourse require a modal logic impress the realist, who is busily occupied in correcting those areas of discourse to make modal logic unnecessary. Each of the two conflicting positions begs the question in attempting to refute the other, for each argues on the basis of premises explicitly denied by the other. It would seem then, that neither position can touch the other at all.

Yet I have characterized functionalism and realism as *conflicting*. Functionalism is not merely *non*-realism; it is *anti*-realism, and the logic of "anti-" is different from the logic of "non". Non-objective art, for instance, is not necessarily anti-objective art. In a treatise on art, objective and non-objective art can be discussed side-by-side, and the contrasts between them made clear. But this is just what could not adequately be done in the case of functionalism and realism. For there is no universe of discourse in which their contrasts can be made clear. Each position claims possession of the only universe of discourse in which comparison can be made at all. Each compares itself to its rival from its own point of view. To ignore this claim is to disqualify oneself from making any comparisons except ones that both of the conflicting positions must disown.

But I still have not shown how positions related in the way that I have maintained that functionalism and realism are related could conflict with each other at all. How can two positions conflict, if neither is capable of regarding the other as a challenge? What I have said about the logic of "anti-" needs supplementation. Let me return to a previous point. The relation between objective and non-objective art is one that it might conceivably be instructive to consider. In other words, a person uncommitted to either type of art could be instructed as to the difference. But it could never be very instructive to consider the difference between functionalism and antifunctionalism in logical theory. For on the face of it, both views are equally arbitrary in a way in which we would not say that objective and non-objective art are arbitrary. Functionalism and realism, to the extent that I have so far described them, seem arbitrary because the sketches I

have given contain no clue as to the *point* of either position. Why on earth should any sane person adopt either one of them? One does not see the *point* of a philosophical position until he grasps the motivation underlying the position. The descriptions of positions in dictionaries of philosophy are usually inadequate because they omit reference to motivation. But, of course, it is insufficient merely to describe the motives from which a philosophical position is taken, since such motives will themselves seem arbitrary. Suppose we are told that functionalism stems from a concern with deduction, and realism from a concern with ontology. We are tempted to ask "What on earth is the point of such odd concerns?" The only way to answer this question is to become involved in the differences between functionalism and realism, instead of being content merely to describe these differences. Being involved in the differences, we shall have to be polemical instead of instructive; we shall have to win over minds committed to an alien position, rather than merely supplying the truth to those who lack it. The unsatisfactoriness of the functionalistic and realistic arguments as I have outlined them so far, arises from the failure of these outlines to suggest the polemical orientation of the arguments. So oriented, the arguments do not sidestep the issues, but meet them and each other head-on.

Realism cannot refute functionalism by pointing to evidence, for the "evidence" it points to is not such as functionalism can even admit to exist. But the realistic *polemic* against functionalism has little in common with the attempted refutation based on alleged evidence. Polemic, rather than making a point that the position under attack must on its own principles ignore, makes a point that the position under attack cannot on its own principles ignore. Thus the full realistic argument against the functionalistic waiver of the law of double negation would *not* appeal to the structure of what is. It would appeal instead to the principles of functionalism itself. But it is necessary to make only a minor change in the wording of the argument in order to implement this major shift in strategy. According to the argument I outlined before, the realist wished to specify precisely what it means to deny a

term so that symbols might express reality without distortion. Instead, the realist might wish to specify the precise meaning of denial in order to make explicit a hitherto implicit presupposition of functionalism. The argument might then run as follows: "Let us grant the functionalist his deviant notions of negation and the queer logics that he develops from these notions. But he will find that he cannot even state these queer logics without making use of normal logic with its customary law of double negation. For if a formula in a queer logic is *not un*provable, it is provable; if a queer system is *not in*consistent, it is consistent; and so on; and these are relationships that the functionalist continually relies on in developing his logic."

The realist can also observe that the functionalist is not interested in *all* deductions; he is interested only in *valid* deductions. Thus the functionalist's investigations do not in fact concern deduction wherever it occurs; they concern deduction only to the extent that it possesses a certain property over and above the mere stepwise movement from premises to conclusion. But the fact that this property itself remains invariant as it applies or fails to apply to every deduction in every domain suggests that the functionalist is actually guided by an ontological consideration over and above all of the peculiar deductive patterns that he investigates. The distinction between a valid and an invalid deduction is, after all, that a valid deduction is a real deduction, while an invalid one is only an apparent one.

Similarly, the functionalist must take the realist seriously if he is to meet the issue. For example, he may try to show that each of the queer logics he recommends is implicit in the realistic defence of an ontologically warranted logic. Validity, invalidity, and inconsistency can, after all, easily be construed as modal properties. The categories of "theorem" and "non-theorem" are not necessarily exhaustive, since in addition to statements that are theorems and statements that are non-theorems there may be formulas that are not statements at all. Indeed, what of statements whose status as theorems or non-theorems cannot be determined? Of course the realist whose

reply consists merely in citing the law of the excluded middle begs the question.

The functionalist can also challenge the realist's intentions in correcting and purifying language. The realist claims that he does this so that the structure of reality can be expressed without distortion. But where does "the structure of reality" really come into the picture at all? If one *defines* "non-P" as "everything but P," then non-non-P is naturally the same as P; but this is certified by the results of deduction, not by ontological insight. It looks as if the real reason why the real reason why the realist wants to refine language is just that he is interested in deduction.

This much will perhaps suffice as an example of the logic of "anti-". I should like to generalize as follows. Suppose there are two philosophical positions A and B such that A is anti-B and B is anti-A. Then there are arguments for A, the propriety of whose premises an advocate of B could not acknowledge; and there are arguments for B, the propriety of whose premises an advocate of A could not acknowledge. Both sets of arguments thus beg the question, and to the extent that the positions rely upon the arguments, they are solipsistically isolated one from the other. Each has reason on its side. But in this situation, not only would it be hopeless for an advocate of A to attempt to convert an advocate of B, or *vice versa*, but also not even the "anti-" would make sense. If A is committed to ignoring all "evidence" for B, and *vice versa*, then there is no genuine opposition. In fact, however, A and B will both be supported by additional arguments, in terms of which there is genuine opposition. The additional arguments for A will consist in attempts to show that B in fact presupposes A; in other words, that implicit in the very formulation of B are presuppositions acceptable to A but inconsistent with the formulation of B. Similarly, among the arguments for B will be found ones claiming to show that A presupposes what is inconsistent with its very formulation as a philosophical position.

The question arises, however, whether these additional arguments can really overcome the solipsistic impasse that seemed to preclude genuine opposition. Suppose an advocate

of A produces a presupposition involved in B that is, in his opinion, inconsistent with the formulation of B. What reason is there to suppose any advocate of B must agree that there is an inconsistency? If he need not agree, the impasse remains.

Of course, the attacker may be mistaken; perhaps what he takes to be an inconsistency between the formulation of B and the presuppositions of B is not really an inconsistency. The question I am now raising is just whether such an attacker must always be mistaken. Is it possible for an inconsistency to develop within a position that "has reason on its side?". If it is *possible*, that will be sufficient to overcome the impasse.

Let me turn the matter the other way round. In point of fact, there will be criticisms of realism to which the realist must reply. It is difficult to see, for example, how he could avoid being obligated to make some sort of reply to the functionalistic objection I have outlined above. To shed light on the necessity of this reply, let us see what the reply is that is necessitated. Consider the functionalist's criticism that modal logics are involved in the development and exposition of classical logic, so that if classical logic is to be the standard of exposition, the realist has failed to meet his own standard. The realist could counter that the *development and exposition* of classical logic is far different from classical logic itself. While classical logic, is, or reflects, the structure of what is, its expositor is thrown upon his inadequate human resources in attempting to express this structure. Who can say but what modal logic simply represents the failure of a finite mind to come to grips with the truth? Perhaps it is just the best we humans can do. Perhaps an omniscient being could do without modal logic.

In replying, the realist has made a distinction. He has been forced to do this by his own failure to meet his own standards of exposition. Thus the distinction he has made was already implicit in his original position, because it was presupposed by the very act of articulating that position. All that the functionalist has done is to have called this fact to the realist's attention. He has not sought to impose anything on the realist. He has merely invited him to overcome disequilibrium by revising his own critical basis. This example,

then, shows how an inconsistency can develop within a position that "has reason on its side," and how the very reason it has on its side may be forced to undergo revision in the effort to overcome the inconsistency.

The problem of this essay can now be stated more specifically than at first. I have said that the position of the realist, at least before he distinguishes classical logic from human thought, is inconsistent and that the functionalist calls this inconsistency to the attention of the realist. Does this not presuppose that the realist and the functionalist both accept the same criterion of inconsistency? If they do, and if the relation between realism and functionalism is really typical of opposed philosophical points of view, then this criterion of inconsistency must be universal, i.e., it must be the same for all philosophers regardless of their philosophical orientations. For any philosopher who wishes to engage in effective criticism of an alien position will have to appeal to this criterion in order to make his criticism clear. But if there were such a universally accepted criterion, philosophical arguments would be subject to non-philosophical criticism —a consequence certainly out of keeping with the view that all philosophical arguments are *ad hominem*. On the other hand, consider the alternative. Suppose there were no criterion of consistency common to two philosophical positions. Then neither could criticize the other. The functionalist, for instance, who thought that he saw an inconsistency in the realist's original position would be met by the reply, "But that is *not* an inconsistency in my terms! If you call it an inconsistency, you are just begging the question!"

In dealing with this problem, I first want to point out that even if conflicting points of view must share a criterion of consistency, it does not follow that any such criterion is universal. All that does follow is that for every pair of conflicting points of view, there is some or other criterion of consistency—not necessarily the same in all cases of conflict.

Also, it is obvious that different criteria of consistency are operative in different conflicts. Sometimes, for example, positions are condemned as inconsistent when contradictory consequences can be drawn from them. Sometimes the alleged

inconsistency is felt to lie in the impossibility of examplifying a position. Neither of these criteria, however, seems particularly relevant to the issue between the functionalist and the realist. The criterion to which appeal seems to be made here is a pragmatic one: no view is consistent if the truth of the view would imply the impossibility of stating the view. I call this criterion pragmatic because of its obvious relationship to what is called the pragmatic paradox; i.e., the paradox of a proposition whose utterance by certain speakers would be self-refuting.

Indeed, the very notion of a universal criterion of consistency immediately destroys itself, because universal criteria of consistency are among the very things that philosophers argue about. Aristotle attempts to defend the law of non-contradiction as a universal criterion of consistency in Book Gamma of the *Metaphysics*. It is highly significant that Aristotle's defense pivots upon precisely the pragmatic considerations adduced in the last paragraph; when someone ignores or denies the law of non-contradiction, his inconsistency must then consist in a systematic incapacity to state his own position. For such a person cannot legitimately say anything, and so *a fortiori* cannot give articulate expression to any particular point of view.

I do not wish to imply, however, that I regard this pragmatic criterion of consistency as having any more of a role in philosophical criticism than does any other criterion. Let us consider the role of the criterion of exemplifiability. By this, I mean the principle that a philosophical position is inconsistent if the contents of the world cannot be interpreted in such a way as to exemplify the position. The use of such a criterion obviously hinges upon an agreement between critic and defendant to the effect that certain interpretations of the contents of the world are precluded. Consider the realistic polemic against the functionalist claim that all deductions are equally worthy of attention. What the realist claims is that this generalization is too inclusive to have any model. It says too much, and therefore says nothing. The most that one could properly say is that all *valid* deductions are equally worthy of attention. This implies that on any possible inter-

pretation, the world contains invalid deductions as well as valid ones; it is impossible to interpret the world as containing valid deductions only. So the issue between the functionalist and the realist presupposes a common ontological commitment. But from this it does not follow that there is any ontological commitment that is presupposed by all philosophical issues. The fact that there is not is immediately proved by the occurrence of arguments over ontological commitments of all kinds. Even the point on which the functionalist and his realistic critic agree has been denied; e.g., by the philosopher for whom all talk is rhetoric and the distinction between valid and invalid talk is an illusion.

What I am proposing then, is to deal with the problem of inconsistency, by asserting that there are a number of criteria of inconsistency no one of which would be appropriate to all philosophical arguments. Any one valid philosophical argument, then, will exhibit an inconsistency determined by the appropriate criteria, i.e. criteria that are or ought to be acknowledged by the partisans of the position under attack. Thus, I have shown how the functionalist attacks the realist in terms of pragmatic criterion, which it is somehow obligatory for the latter to accept; and when the realist attacks the functionalist, he employes an ontological criterion which is somehow implicit in the functionalist's own position. I do not at the moment know how to describe the mechanism which I have indicated by "somehow". In fact, it seems odd, on the face of it, that the realist should be committed to a pragmatic criterion, and the functionalist to an ontological one; one might have supposed that it would be the other way around. But on second thoughts, dialectical mechanisms do often operate in this way. In any event, this seems to me to be a way of defending the purely *ad hominem* character of philosophical argument against the objection that inconsistency is a matter of fact.

The Pennsylvania State University.

VOL. LXXVIII. No. 309] [January, 1969

MIND
A QUARTERLY REVIEW
OF
PSYCHOLOGY AND PHILOSOPHY

I.—ON SELF-REFERENCE AND A PUZZLE IN CONSTITUTIONAL LAW

BY ALF ROSS

1. *Introduction*

IN an early work (*Theorie der Rechtsquellen*,[1] 1929) and also in subsequent publications (*On Law and Justice*,[2] 1958, Danish original 1953; *Dansk Statsforfatningsret*[3] (Danish Constitutional Law, 1959-60) I have already pointed out a puzzle involved in the idea of a constitution comprising rules for its own amendment, if these rules are regarded as part of the constitution and as subject to the amendment procedure which they themselves lay down. To understand how this puzzle arises, it is necessary to make a few preliminary observations.

It is a fundamental trait of a legal order that most of its constituent rules are created through enactment, that is through a human decision in accordance with other legal rules called *rules of competence*. A rule of competence prescribes the (necessary and sufficient) conditions under which an enactment shall be valid, that is have force of law. Typically these conditions may be divided into three sets: (1) one pointing out the human being or beings qualified to undertake the enactment; (2) another describing the procedure of the enactment; and (3) a third limiting the subject matter of the rule to be enacted by these persons and by this procedure. Thus, for example, a will creates valid rules of succession, if it has been drawn up by the right person in the right way and contains dispositions in accordance with the law

[1] Chapter xiv. [2] § 16. [3] §§ 41 and 46.

of succession. Or, again, valid statutory law is created, if an act has been passed by the qualified persons (in England the Queen and the lawful members of Parliament) according to a correct procedure (concerning the summoning, prorogation and dissolution of Parliament, voting methods, etc.) and it deals with matters within the scope of its constitutional powers.

We shall say that any rule of competence constitutes an *authority*. Sometimes we shall use this term in its current unsophisticated use as designating the person or body of persons invested with the law-enacting power, *e.g.* Parliament, the Government, a Secretary of State, a local authority, or a regulatory commission. Generally, however, we shall use the term as a symbol for the totality of conditions determining the law-enacting process. In particular, we shall say that a norm of competence constitutes an authority defined by the sum of the conditions necessary and sufficient for the enactment of legal rules. The authority in this sense, we could also say, is a personification of the whole law-enacting process defined through the rules concerning personal, procedural, and material competence.

Enacted law is also called *written* law. Not all law is written law, cases in point being customary law and the law emerging from the practice of the courts.

A norm of competence (constituting an authority) may itself be written law, that is, enacted by another authority. One authority may thus be constituted by another. As the legal validity of the first is derived from the second, it is natural to consider the constituting authority as of a higher type, or belonging to a higher level than the constituted authority. Two authorities constituted by the same superior authority are on the same level. Two authorities constituted by two different authorities on the same level are also on the same level, and so on. In this way there arises a complicated system of authorities on various levels.

Let us consider a certain authority A, constituted in a set of rules of competence C_1. Now C_1 must either be or not be enacted by a higher authority A_2 constituted in a set of rules of competence C_2. If it is so enacted, what was said about C_1 holds also for C_2 : it must either be or not be enacted by a higher authority A_3, and so on. Since the series of authorities cannot be infinite, the inevitable conclusion follows that there must be a highest authority whose competence is not derived from any other authority. The line of thought can be illustrated by the following figure :

A_1 is constituted by C_1 ; C_1 is enacted by A_2
A_2 is constituted by C_2 ; C_2 is enacted by A_3
A_3 is constituted by C_3 ; C_3 is not enacted by any other authority.

A_3, then, is the *highest authority* of the system and C_3 its *basic norm.*[1] But what does it mean to say that its constituting set of norms of competence C_3 is not enacted by any other authority ? How, in that case, could the legal existence of C_3 be established ? Two, and only two, answers seem possible. But both seem unacceptable. That is the puzzle.

The two possible answers are :

(1) C_3 is enacted law ; as it is not enacted by any other authority this means that it is enacted by A_3 itself ;

(2) C_3 is not enacted law ; this means that its legal validity cannot be derived from the validity of any other norm, but is an original fact, a presupposition for the validity of any other norm of the system.

Before I go on to explain why both answers seem to be unacceptable, it is appropriate to state how the problem arises in actual constitutional law.

The essential task of any constitution is to establish a legislative authority. If we identify this with A_2 in the above account, the constitution includes the norms C_2, determining the competence of the legislator. If now the constitution contains rules for its own amendment, these rules (C_3) determine another law-enacting process and constitute another and higher authority, usually called the constituent authority or power, corresponding with A_3 in the figure above. If the constitution knows no higher authority competent to amend the amendment rules, then A_3 is the highest authority of the system and C_3 its basic norm. Thus for example, the highest authority in Denmark is the constituent power established by the rules of amendment in art. 88 of the constitution of 1953, and art. 88 itself is the basic norm of the Danish legal order.

The question now arises : How can art. 88, that is the rules C_3 constituting the highest authority A_3, itself be amended ? And the two answers are :

[1] The idea of a legal order as a systematic unity implies that there must be either only one highest authority or a plurality of co-ordinated authorities at the highest level.

(1) art. 88 may be amended according to its own rules, that is by enactment by the constituent power A_3 constituted by art. 88 itself ; or

(2) the legal validity of art. 88 being an original fact, underived from the validity of any other norm, there is no legal procedure according to which art. 88 may be amended. This does not mean that it is unchangeable. For just as one legal custom may be supplanted by another legal custom, so one basic norm may be supplanted by another. But the transition in both cases is not the outcome of a legal process ; it is a fact, the sociopsychological fact that the community now accepts another custom as law or another basic norm as the corner-stone of its legal order. It is true, of course, that as an historical fact a proposal for amending art. 88 may be passed by a process in accordance with the prescriptions of art. 88, with the result that the amended art. 88—let us call it art. 88′—is generally accepted as law. In that case, however, the legal validity of art. 88′ still cannot be derived from art. 88 and the amendment procedure, but emanates directly from the fact that it is accepted by the community as law. We can also put it in this way : all legal enactment must be intra-systematic, that is, go on inside a given legal order in accordance with a rule of competence belonging to the order. The transition from one basic norm to another, that is from one system to another system, therefore cannot be the outcome of an enactment, but must be an extra-systematic fact amounting to the founding of a new system that replaces the old one.

It is now time to explain why neither of the only two possible answers seems to be acceptable.

The first answer—that the basic norm is enacted by the highest authority, or, what is the same, that the basic norm may be amended in accordance with itself—seems unacceptable, because it runs contrary to an alleged logical theorem according to which self-referring sentences are devoid of meaning. I shall return later to this disputed theorem and confine myself here to pointing out that the assumption that the basic norm may be amended in accordance with its own rules involves contradictions.

Let us remember first that, when we consider a certain rule to have been validly created through enactment by a certain authority, our reasoning takes the shape of an inference of the pattern $[(p \rightarrow q) \,\&\, p] \rightarrow q$ in which the first premiss is a statement of the conditions under which, according to the norm of competence, a valid norm is created, the second premiss the statement that those conditions are fulfilled, and the conclusion the statement that a valid norm has been created. For example :

A norm is valid, when created in accordance with conditions C_1, C_2, and C_3 ;
The norm N has been created in accordance with the conditions C_1, C_2, and C_3 ;
$\therefore$ The norm N is valid.[1]

Now, if we suppose art. 88 to be amended according to its own rules with the result that it is replaced by art. 88′ (with a content contrary to that of art. 88) the validity of art. 88′ is based on an inference of the following pattern :

art. 88 : The constitution may be amended by a process in accordance with conditions C_1, C_2, and C_3, and only by this process ;

art. 88′ (stating that the constitution may be amended by a process, in accordance with conditions C'_1, C'_2, and C'_3) has been created in accordance with conditions C_1, C_2, and C_3

$\therefore$ art. 88′ is valid, that is, the constitution may be amended by a process in accordance with conditions C'_1, C'_2, and C'_3 and only by this process.

As the meaning of art. 88 is to indicate the *only way* in which the constitution may be amended, this is an inference in which the conclusion contradicts one of the premisses, which is a logical absurdity.

Although this is sufficient proof of the impossibility of amending a basic norm in accordance with its own rules, the reasoning may become more convincing, if we consider cases less complicated than art. 88 of the Danish Constitution.

Suppose the basic norm of a paternal relationship to be that the son S is in every respect subject to the will of his father F. Now, if F tells S not to listen to him any more, but to have his own way, the emancipation of S cannot be based on an inference from the basic norm, because in such an inference the conclusion (the emancipation) would contradict the first premiss (the basic norm of parental rule). If S really makes autonomous decisions, because (on the grounds that) F has told him to do so, he still accepts the parental authority and is not really emancipated ; a new parental ruling could destroy his independence. On the

[1] I do not take up the point that this inference, in spite of the indicative expressions used, is really a directive (deontic) inference. The first premiss is really a norm prescribing an obligation to obey the norms created in the way indicated, and the conclusion is also a directive.

other hand, if *S* really and unconditionally realizes autonomy as his new basic norm, this cannot be derived from the old norm of parental authority. The new norm holds, not because it has been commanded by *F*, but because of the psychological fact that it has been accepted by *S*.

Or, let us suppose a basic norm granting absolute power to a monarch. Now if, as happened in a number of states in the nineteenth century, the absolute ruler on his own authority grants his people a " free constitution ", the legal significance of this act may be interpreted in two different ways. If the new constitution is considered as being granted and having validity by virtue of the absolute power of the monarch, absolute rule is still the basic norm and the freedom may at any time be revoked by the monarch. This was undoubtedly the case, when in 1831 the Provincial Estates were introduced in Denmark. The new constitution, however, could also be understood as a definite termination of absolutism. In that case its validity could not be derived from the Royal Act, but would have to be considered as a new historico-political fact. Whether the one or the other interpretation is realized, in fact, is neither a legal nor a logical, but a political question dependent on what ideology actually dominates the minds of people.

Similarly, if we assume a basic norm that requires a majority of 60 per cent. for amendment of the constitution. If such a majority decides that in future the majority is to be 70 per cent., this new basic norm cannot be derived from the old one. And so also with respect to the actual art. 88 whose rules, however, are rather more complicated.

The second answer—that the basic norm is not enacted law and therefore cannot be amended by any legal process—seems to fare no better. It is contrary to the undeniable fact that people (in Denmark, as elsewhere) think and act as if the basic norm (art. 88) may be amended in accordance with its own rules. There is no doubt that any attempt to change art. 88 in any other way would be considered illegal by the people, the leading politicians and the courts. So it must be accepted as Danish constitutional law that art. 88 may be amended by the legal process described in this article itself. I have always accepted this view though with the addendum that in fact people are dominated by ideas that cannot be expressed rationally, but only in magical terms. That is, amendment of the constitution is conceived as an act of magic in which he who possesses the " power " conjures its transference to another. In magic the sequence in time is decisive ; the conjuration terminates the original power. A logical inference, on

the other hand, knows of no sequence in time; the conclusion cannot destroy the validity of any premiss " for the future ".

Today I cannot uphold this explanation. If people do in fact act on a certain understanding of the meaning of art. 88, this understanding *must* be expressible in rational terms. This means that the basic dilemma is still unsolved. Faced with the exclusive alternative as to whether the basic norm is or is not enacted law, that is, as to whether it can or cannot be amended by a legal process according to rules of competence (amendment rules), we have to admit that neither of the two possible answers seems acceptable. This is the challenging puzzle. My earlier discussions of these problems have elicted a number of comments and attempts to deny or do away with the difficulties inherent in the self-reference of art. 88—in my view unsuccessfully.[1] The aim of this study is to find a solution of the puzzle. To this end I shall first take up the general problem of self-referring sentences and then apply the outcome of this inquiry to the constitutional paradox.

First Part: Self-Reference

2

Russell's theory of types is based on the idea that a number of well-known paradoxes result from a certain kind of vicious circle. The vicious circles in question arise from supposing that a collection of objects may contain members which can only be defined by means of the collection as a whole. The first example given by Russell is a proposition stating that " all propositions are either true or false ". It would seem, he says, that such a statement could not be legitimate unless " all propositions " referred to some already definite collection, which it cannot do if new propositions are created by statements about " all propositions ". We shall therefore have to say that statements about " all propositions " are meaningless. To avoid the illegitimate totalities which generate the paradoxes Russell sets up the " vicious-circle principle " : " whatever involves *all* of a collection must not be one of the collection ".[2]

It is easy to see that to apply this principle to propositions is

[1] Max Sørensen in *Juristen*, 1959, pp. 446 ff.; Poul Andersen, *loc. cit.* 1960, pp. 507 ff.; Niels Egmont Christensen, *loc. cit.* 1960, pp. 231 ff.; Torkel Opsahl in *Tidsskrift for Rettsvitenskap*, 1962, pp. 314 ff.; Ole Krarup i *Ugeskrift for Retsvæsen*, 1967, pp. 10 ff.

[2] Bertrand Russell and Alfred North Whitehead *Principia Mathematica* (2nd edn., 1960), vol. i, p. 37.

to rule out self-reference or reflexivity as illegitimate. The scheme " all propositions are either true or false " is illegitimate, because it refers not merely to all other propositions, but also to itself. After having dealt with a number of paradoxes, Russell himself draws this conclusion. He says " In all the above contradictions (which are merely selections from an indefinite number) there is a common characteristic, which we may describe as self-reference or reflexiveness. The remark of Epimenides must include itself in its own scope. If *all* classes, provided they are not members of themselves, are members of w, this must also apply to w; and similarly for the analogous relational contradiction." [1]

Russell's theory of logical types recommends itself, as the author says, in the first instance by its ability to solve certain contradictions. But the theory in question, he adds, is not wholly dependent upon this indirect recommendation; it has also a certain consonance with common sense which makes it inherently credible.

Now, if one accepts Russell's theory of types, one has implicitly accepted also the ban against reflexive propositions. But the contrary does not hold: the complicated theory of types is not a necessary presupposition for the theorem that self-referring propositions are devoid of meaning.

This is stressed by Jørgen Jørgensen. True, he says, the theory of types has, if tenable, shown how the paradoxes can be avoided. But this result can be obtained in a much simpler way which has the further advantage over the theory of types that it is able to explain how the paradoxes arise. He says

> My point is that the expression " This sentence is false " is not a sentence at all in any logical sense. If the expression mentioned is considered a sentence, and the word " false " is considered the logical predicate of this sentence, then the logical subject cannot be the whole expression, but at most either the words " This sentence " or the designation of these words. In the first case the whole expression is meaningless, because " false " is not the kind of predicate which can be meaningfully ascribed to descriptions as " This sentence ". And in the latter case the whole expression is meaningless, because the description " This sentence " has no object, there being no sentence to which these words can refer. In neither case can any conclusions be drawn from the expression mentioned, and no paradoxes emerge.[2]

In following Jørgensen, I have tried to make his reasoning more convincing by a method of transcription. It seems to me an

[1] *Op. cit.* p. 61.

[2] Jørgen Jørgensen, " Some Reflections on Reflexivity ", MIND, lxii (1953), 289 ff., 290.

obviously legitimate claim that if a description contains a referring phrase, this phrase must be replaceable by the object to which it refers if the description is to have any meaning at all. For instance, if one says " This man is wise ", the referring phrase " this man " must be replaceable by " Joe Smith " or " John Brown ", or the name of whoever is intended, if the sentence is to have any meaning at all. If someone said to me " This man is wise " and upon asking him " Who do you mean ? " I got the answer " Well, I can't say—I just mean ' This man is wise ' ", I would certainly feel justified in ignoring his utterance as plain nonsense.

Therefore, if the sentence " This man is wise " is to have any meaning, it must be transcribable in the following way :

" This man (that is, the man John Brown) is wise."

It is easy to see that if we try to transcribe the sentence " This proposition is false " in a similar way, we lose ourselves in an infinite regress and never get an answer to the legitimate question " Which proposition ? " In a first attempt we would get this transcription :

" This proposition (that is, the proposition ' This proposition is false ') is false."

But because the transcription itself contains a referring phrase, a new transcription of this transcription is required, and so on *ad infinitum* : one is never told what proposition is subject to the qualification of being false. Consider a traffic sign announcing the nearest way to Oxford in the following terms : " The nearest way to Oxford is as announced on this board."

We are similarly frustrated, if we try to transcribe a partly self-referring sentence, such as art. 88 of the Danish Constitution. If we call the process of amendment prescribed by this article P, its meaning may be transcribed thus :

Art. 88 =
Art. 1 (stating that . . .) is amendable by process P ;
Art. 2 (stating that . . .) is amendable by process P ;
.
.
.
Art. 88 (stating that . . .) is amendable by process P ;

In each case the parenthesis is to be filled in with the prescription contained in the article in question. When we arrive at art. 88 this means that we have to begin again with art. 1 and go on to art. 88, and then begin again with art. 1, and so on *ad infinitum.*

3

These arguments support, I believe, the presumption that there must be something wrong with self-referring sentences. However, since prominent philosophers (such as K. R. Popper, H. L. A. Hart, Ivar Segelberg, Niels Egmont Christensen, and K. Grue-Sørensen)[1] have defended such sentences or, at any rate, certain kinds of them, it is necessary to look deeper into the matter. First of all it seems that we should state the problem and define self-reference more precisely. To this end it is appropriate to clarify certain distinctions.[2]

A speech-act is essentially a phonetic act, *i.e.* the production of a sequence of sounds (or symbols for sounds). These sounds are psycho-physical phenomena. As a physical phenomenon, the phonetic act may produce effects quite outside the process of communication, *e.g.* when shouting causes an avalanche in the Alps.

Not every production of a sequence of sounds, however, is a speech-act. The phonetic act must further possess a structure which accords with the syntactical rules of the language concerned, *i.e.* the rules governing the ways in which the linguistic elements may be combined into compound wholes. These rules include especially those which govern the structure of sentences, *i.e.* the grammatical syntax according to which such a word sequence as " That failed of boys yesterday because " does not count as a sentence.

Not every grammatically correct sentence, however, constitutes a speech-act. For this it is further required that the sentence possess meaning. First of all, a grammatically well-formed sentence may be rejected as a candidate for a speech-act, if it is contrary to the syntax of formal logic as, for example the sentence, " It is raining and it is not raining ". Besides formal logic there is also a semantic logic that rules out such sentences which accord with the requirements of formal logic as " Five per cent. of the prime numbers, having as their father the concept of temperature and as their mother the number five, die, within a period of three years plus five pounds plus seven inches after their

[1] Karl R. Popper, " Self-Reference and Meaning ", MIND, lxiii (1954), 162 ff.; H. L. A. Hart, " Self-Referring Laws ", *Festskrift till Karl Olivecrona* (1964), pp. 307 ff.; Ivar Segelberg, " Bemerkungen zu einigen logischen Paradoxien ", *Theoria* (1943), pp. 157; Niels Egmont Christensen, " Er grundlovens § 88 en del af grundloven ?" (Is Art. 88 of the Constitution Part of the Constitution ?), *Juristen*, 1960, pp. 231 ff.; K. Grue-Sørensen, *Studier over Refleksivitet* (Studies in Reflexivity) 1950.

[2] *Cf.* my book, *Directives and Norms* (1968), pp. 3 ff.

birth, of either typhoid fever or the square root of a democratic constitution ".[1] Semantic logic has been but little investigated. The question with which we are concerned is whether or not a ban against self-referring propositions makes out a rule of semantic logic that debars such sentences even though they do not, at least openly, conflict with the rules of formal logic.

Corresponding to this triad, we must distinguish between the speech-act as (1) a *sound-sequence*, (2) a *sentence*, and (3) the owner of a *meaning*. The meaning content may be either a *proposition* (when the speech-act occurs in indicative speech) or a *directive* (when the speech-act occurs in directive speech).

As we have said, the question is whether sentences which suffer from the vice of reflexivity must be rejected as meaningless. We must now try to define the defect characterizing this group of sentences more precisely.

Self-reference is usually taken as the referring of a sentence to itself. But this is ambiguous. It must be clarified *where* the reference occurs and *what* its object is.

Now, since " reference " is a semantic concept, the speech-act itself either as a sound sequence or as a grammatical construction, a sentence, is unable to refer to anything at all. The reference occurs in the meaning. If I say, for example, " This man is wise ", the reference is in the proposition, I express, in particular in the meaning of the phrase " this man ". The reference is such that the meaning is incomplete or non-existent until the reference has been filled in, that is, until the man intended has been pointed out.

It seems to me a reasonable hypothesis that the vice of self-reference occurs when one tries to express in a sentence a meaning which refers to the meaning of the same sentence. In that case no filling in is possible, one remains empty-handed, while there is no defect in a sentence expressing a meaning that refers either to the sentence itself as a grammatical construction or to the speech-act as a sound sequence. Some reflections corroborate this view. Consider the following sentences :

(1) The sounds emanating from my mouth between t_0 and t_1 are so faint that they cannot be perceived at a distance of five yards ;
(Between t_0 and t_1 I actually pronounce (1)).
(2) The sounds I am now pronouncing are so faint . . . ;
(3) What I now say is said so softly that . . . ;

[1] Borrowed from Rudolf Carnap, *Einführung in die symbolische Logik* (1954), p. 76.

(4) The sentences pronounced by me between t_0 and t_1 are well-formed sentences in the English language ;
(Between t_0 and t_1 I actually pronounce (4)).
(5) The sentence I am pronouncing is a well-formed sentence in the English language ;
(6) This sentence is a well-formed sentence in the English language.

It seems to me that (1) and (4) are indisputably unobjectionable constructions. Their meaning is clear. The reference is to a clearly identifiable object, a certain sequence of sounds or a certain grammatical construction, a sentence. However, the same is the case with regard to (2), (3), (5), and (6). The only difference between these sentences and (1) and (4) is that the reference is more loosely expressed in (2), (3), (5), and (6). If we transcribe the referring phrases, occurring in (2), (3), (5), and (6), we do not lose ourselves in an infinite regress. The reference in these cases is either to a sound sequence or to a sentence, not to a proposition. As neither a sound sequence nor a sentence as such refers to anything, once the first transcription has taken place, there is nothing more to transcribe. The kind of self-reference that occurs in (3) and (6) is harmless. Let us call it *spurious self-reference* in contrast to the *genuine self-reference* which leaves a sentence without meaning.

4

In this and the following sections I am going to take up, one by one, the arguments which have been advanced in the literature against the theorem that (genuine) self-reference deprives a sentence of meaning.

First it should be noted that most of the instances given to demonstrate self-referring sentences with good meaning are instances of spurious self-reference and therefore without any force as arguments against the well-formulated theorem. This holds with respect to the following examples :

" I am now speaking so softly that dear old Socrates cannot make out what I am saying " (Popper) ;

" This statement is written in English " (Hart, Grue-Sørensen) ;

" I am whispering this statement " (Hart).

But we cannot deal in the same way with a further example given by Popper, namely the sentence :

" What I am now saying is meaningful "

which Popper believes he can prove to be true and meaningful. Here we have a case of genuine self-reference, and if Popper were really able to give such a proof, the self-reference theorem would either be false or have to be modified.

But Popper's alleged proof is not conclusive. He believes it to be demonstrable by means of the *reductio ad absurdum* method. " I assume for the purpose of the *reductio* ", he says, " the truth of the negation of [my] theorem, that is, the truth of the assertion : ' What I am now saying is meaningless '. If this assertion is true, it must be, clearly, meaningful. Thus the assumption that it is true is absurd ; which proves [my] theorem." [1]

However, this " proof " is nothing but a vicious circle. By ascribing hypothetical truth-values to the disputed sentence it is already assumed that the sentence has meaning—which is exactly what should be proved. Although Popper is right in so far as he maintains that it is absurd to assume the assertion " What I am now saying is meaningless " to be true, it does not follow from this that the assertion must be false. There is also the possibility that it is neither true nor false, because it is meaningless.[2]

5

Popper does not believe that paradoxes such as the *Liar* (who says " What I am now saying is untrue ") can be solved by dwelling on the impossibility of self-referring assertions. For even if direct self-reference was impossible, or meaningless, indirect self-reference is still possible and certainly quite common. As an example, Popper gives the following dialogue between Socrates and Theaetetus :

> *Th.* The very next question which I am going to ask you is an extraordinary one, although expressed in perfectly ordinary language.
> *S.* There is no need to warn me : I am all ears.
> *Th.* What did I say between your last two interruptions, Socrates ?
> *S.* You said : " The very next question which I am going to ask you is an extraordinary one, although expressed in perfectly ordinary language."

The point is that although Theaetetus' first and second utterance refer to each other, Socrates, in spite of this indirect

[1] *Op. cit.* (above, p. 10, note 1), p. 165.
[2] The same objection is pertinent to his discussion (on pp. 166-167) of the question : " Is the question I am now asking you meaningful or meaningless ?"

self-reference, understands them perfectly well, which indicates that they cannot be meaningless.

I agree with Popper in the belief that it cannot, in principle, make any difference whether self-reference appears directly or indirectly. Self-reference is indirect, when S_1 (statement number one) refers to S_2, with the result that the meaning of S_1 is incomplete, until it is filled in with the meaning of S_2 and *vice versa*. Thus, for example, if A tells you that the nearest way to Oxford is as B will tell you, and B tells you that the nearest way is as A has told you, this information is no more illuminating than the traffic sign mentioned above in section 2. The *Liar*, as Popper points out, can just as well be stated indirectly, as for example in " The next assertion I am going to make is a true one " followed by the " The last assertion I made was untrue ". Therefore if indirect self-reference is, as Popper says, quite a common thing, direct self-reference must also be accepted as legitimate.[1]

Popper's example, however, is not a case of genuine self-reference ; it does not involve that circularity that deprives the connected utterances of meaning. Since the introduction of a question in the example complicates the issue, I prefer to restate the example using statements only. This transcription in no way affects the illustrative force of the example. We get :

> S_1 : The statement I am going to make in a moment is an extraordinary one ;
> S_2 : A moment ago I said : " The statement I am going to make in a moment is an extraordinary one."

It is true that S_1 refers to S_2 in the sense that the meaning of S_1 is incomplete, until one knows the meaning of S_2. But the same does not hold with respect to S_2. Its meaning is complete without filling in. It does not refer to S_1 in the sense in which S_1 refers to S_2. S_2 *speaks about* S_1, but does not incorporate the meaning of S_1 as part of the meaning of S_2. S_2 states a definite historical fact, namely the fact that a moment ago I pronounced a certain *sentence* whatever its meaning. So no circularity and no self-reference arise. The conclusion is that however commonplace a connection between two sentences as demonstrated in Popper's example may be, it has nothing to do with genuine self-reference.

6

Popper also says that he is inclined to argue that if the meaning of an utterance can be understood then the utterance *has* a meaning.[2] This, together with the assumption that Socrates

[1] *Op. cit.* pp. 162-164. [2] *Op. cit.* p. 166.

was able to understand Theaetetus' theorem, " What I am now saying is meaningful ", is taken as another proof that this theorem, in spite of its genuine self-reference, has a meaning.

This line of reasoning raises the delicate question as to what it means to assert, and how it is ascertained, that Socrates has really understood anything. Popper says nothing about that. Clearly, he relies blindly on Socrates' testimony. But I find that for my own part I cannot do that. I must strictly deny that I understand anything, when I hear the sentence " What I am now saying is meaningful ". If someone really said this, I would wait for what he was going to say. If he said nothing more, I would consider his utterance an unfinished message without meaning. But, of course, this is only my testimony against that of Socrates. I wonder whether Socrates would insist also that he was able to understand a self-referring directive such as " Obey this directive ", or the traffic sign mentioned above in section 2. At least I feel sure that whatever " understanding " he had would not help him to perform any act that satisfied the " directive " nor help him to find his way to Oxford.

7

In Hart and Christensen we meet the idea that although a totally self-referring proposition—that is a proposition that refers to itself and nothing else—is devoid of meaning, this is not the case if the proposition is only partly self-referential, that is, if it refers to itself as part of a greater whole. They both agree that, *e.g.* the sentence " All sentences on this page (including this one) express true propositions " is meaningful. Christensen further assumes that the sentence " All sentences on this page (including this one) express false propositions " is meaningful and necessarily false.[1] Hart reserves his opinion with regard to this latter sentence.[2]

It does not appear plausible that partial self-reference should fare better than total self-reference. It seems reasonable to suppose that if S refers to A, B, C, and S, and assuming genuine self-reference, S is meaningful in so far, as it refers to A, B, and C, and meaningless in so far, as it refers to itself. I cannot see that either of the authors has given reasons for his view to the contrary. My impression is that their view is motivated exclusively by the fact that the assumption of the meaningfulness of such sentences does not give rise to paradoxes. Paradox,

[1] *Op. cit.* (above, p. 10, note 1), p. 236.
[2] *Op. cit.* (above, p. 10, note 1), p. 316.

however, does not necessarily accompany self-reference. Although the sentence " This sentence is true " does not give rise to paradoxes, it is still just as meaningless as the paradox-creating *Liar*.

The proof given by Christensen that the sentence " All sentences on this page are false " necessarily must be false and therefore meaningful, is of the same *reductio ad absurdum* pattern as Popper's proof mentioned above in section 4, and inconclusive for the reasons given there.

The partial emptiness of S is seen if we apply the method of transcription mentioned above in section 2 :

S : all sentences on this page are true
= the sentences A, B, C, and S are all true
= A (stating that . . .) is true ;
B (stating that . . .) is true ;
C (stating that . . .) is true ;
S (stating that . . .) is true.

In order to fill in this last parenthesis we have to start once more from the beginning and so on *ad infinitum* ; it is impossible to find any other meaning than the assertions that A, B, and C are true.

One could try to save the partially self-referring sentences by the following artifice. For S we substitute a number of other expressions such that (1) self-reference is avoided, and (2) it is not unreasonable to consider the meaning of the substituted expressions as identical with the intended meaning of S. Those two conditions seem to be satisfied, if S is replaced by a finite series of sentences plus a sentence about the series :

$S = S_1$: A, B, and C are true
S_2 : S_1 is true
S_3 : S_2 is true
.
.
.
S_n : S_{n-1} is true
$+S'$: S_n is true for any value of n.

It can easily be shown that a similar substitution is impossible if S = " This sentence is true " or S = " All sentences on this page are false ".

Even if this substitution is accepted, we have not shown that the partly self-referring sentence has any meaning beyond the part of it that is not self-referring. For S_n = " It is true that

it is true that it is true . . . that A, B, and C are true " says not one jot more than S_1, namely that A, B, and C are true.[1]

8

In section 2 I mentioned the method of transcription by means of which I have tried to make the emptiness of self-referring sentences obvious. Christensen maintains that my reasoning involves a circle : to insist on a transcription of reflexive sentences presupposes what should be proved, namely that they are unintelligible, as they stand in themselves.

This objection is unfounded. The demand for a transcription is not directed especially towards *self*-referring sentences, but towards any reference.

It seems to me obvious that whenever a referring expression is used, it must be legitimate to ask to what it refers—and to transcribe the reference accordingly. When someone says " This man is wise " it must be legitimate to ask " What man ? " ; and when someone says " This proposition is true ", it must be legitimate to ask " What proposition ? "

9

Summing up this part of the paper, we must conclude that none of the objections to the logical condemnation of reflexive sentences has proved itself tenable. It must still be considered a plausible thesis that genuine self-reference, whether direct or indirect, whether total or partial, deprives the sentence of meaning as far as self-reference is involved.

Second Part : Solution of the Constitutional Puzzle

10

In the introduction to this paper I explained how the constitutional puzzle arises : it seems impossible to give an acceptable answer to the question about the establishment and amendment of the basic norm of a legal system, *e.g.* art. 88 of the Danish Constitution. The two possible candidates for an answer are :

[1] This gives rise to a question with far-reaching implications, namely whether the sentence " p is true " expresses any proposition, that is whether it itself can be true or false. I am inclined to deny this and to believe that " p is true " expresses an attitude-deciding act, the acceptance of p, which may be well-founded or ill-founded, upheld or withdrawn, but which cannot be either true or false.

(1) that art. 88 may be amended in the legal process, it itself prescribes, that is, by the highest authority which the article itself constitutes. This answer is rejected for two reasons :

(*a*) because it involves genuine, partial self-reference which must be ruled out as a logical absurdity ; and

(*b*) because it involves the assumption of an inference in which the conclusion is contrary to one of the premisses, which again is a logical absurdity ;

(2) that art. 88 cannot be amended in any legal process, but only as the outcome of the socio-psychological fact that the society actually accepts another basic norm as the corner-stone of its legal order. This answer is rejected, because it is contrary to obvious facts.

We are now going to consider various proposed attempts to solve this puzzle. They are all designed to favour the first answer by contesting the two mentioned objections against it. (These objections must of course both be refuted, if the vindication is to be deemed successful.)

In the first part of this paper I have considered the various attempts to defend self-reference, but have rejected them all in so far as concerns genuine self-reference. If this is correct, it is in itself sufficient to refute any attempt to argue for the first answer. And as nobody has ventured to defend the second answer, this means that the paradox remains unsolved.

In particular I want to note that the method of substitution which we mentioned in section 7 is unable to save art. 88 from the defects of genuine, partial self-reference. If we try to apply this method, we get :

Art. 88 =

Art. 88_1 : The articles 1 to 87 of the Constitution are amendable by process P ;

Art. 88_2 : Art. 88_1 is amendable by process P ;

Art. 88_3 : Art. 88_2 is amendable by process P ;

.

.

.

Art. 88_n : Art. 88_{n-1} is amendable by process P ; in combination with the rule that art. 88_n is valid for any value of n.

If we now suppose that a new norm for constitutional amendment (let us call it art. 88′), which substitutes process Q for process P, is itself established in process P, we get, whatever the value of n, the following series :

Art. 88′ =
Art. $88'_1$: The articles 1 to 87 of the Constitution are amendable by process *Q*;
Art. $88'_2$: Art. $88'_1$ is amendable by process *Q*;
Art. $88'_3$: Art. $88'_2$ is amendable by process *Q*;
.
.
.
Art. $88'_n$: Art. $88'_{n-1}$ is amendable by process *P*.

The point is that, whatever the value of *n*, art. 88_n as the ultimate basis for the enactment must itself remain unchanged. This implies that the basic norm remains unchanged and shows that it is impossible to remove reflexivity by this method. This would require that an amendment of art. 88_{n-1} in accordance with art. 88_n at the same time amends also art. 88_n itself. But that precisely implies the self-reference which should be avoided.

Nevertheless, in case anybody should not be convinced of the soundness and force of my arguments against genuine reflexivity, or of their applicability to the problem of constitutional amendment,[1] I shall take up different arguments by means of which various authors have tried to meet my second objection to the first answer, namely that it involves a contradiction within an inference.

11

Hart and Christensen both pose the question whether it is correct to say that the amendment presupposes an inference in which the amended article is the conclusion, though neither of them stresses this point.

In my opinion there can be no doubt that in legal as well as popular reasoning the legality of the amendment procedure and the validity of the amended article are based on an inference: *As* art. 88 is valid constitutional law and *as* the amendment conditions prescribed in this article are fulfilled, so it *follows* that art. 88′ is now valid constitutional law. I refer to the introduction where it was mentioned that an inference of this

[1] Hart, *op. cit.* (above, p. 10, note 1), p. 315, raises the question whether the principles of reflexivity which may hold in relation to propositions are applicable also to norms, especially legal rules. I do not see why they should not. The rule against self-reference is concerned with the meaning of a speech-act and is independent of whether the meaning content is used to state how the world *is* or to prescribe how it *ought to be*. The directive: "Do not obey this command" is just as meaningless as the *Liar*.

type occurs whenever a rule is enacted in accordance with a norm of competence. The question we are dealing with is precisely whether or not this line of reasoning is logically sound, when applied to the basic norm.

12

Christensen argues that it is not excluded that in an inference the conclusion may contradict the premisses, and gives as an example the classical proof of the irrationality of $\sqrt{2}$.[1] This requires no comments and must have been written in a moment of distraction.

13

The main argument relied on by all my critics is that, when the sequence of time is taken into account, there is no contradiction between art. 88 and art. 88′ : art. 88 ceases to be valid law from the moment, art. 88′ comes into force. This argument, however, confounds legal with logical contradiction. There is no contradiction *in law*, because art. 88′ supersedes art. 88. But why does art. 88′ supersede art. 88 ? Precisely because art. 88′ logically, that is, corresponding to its meaning content, contradicts art. 88. This follows from the well-known *lex posterior* principle according to which in case of conflict between two equivalent norms (*i.e.* norms at the same level of the norm hierarchy) the later precedes the earlier one.

It is understandable that lawyers accustomed to the idea that the legal order is without contradictions (because of the accepted principles for the solution of conflicts between immediately contradictory norms) could fall into this trap. It is less understandable that the philosophers taking part in the discussion should have done likewise.

The contradiction cannot be avoided by building the *lex posterior* principle into art. 88 itself. On this interpretation the meaning of the article could be stated as :

Art. 88 : The rules of the Constitution are amendable by process P and only by this process, until by this process it is decided otherwise.

And the meaning of art. 88′ would be :

Art. 88′ : The rules of the Constitution are amendable by process Q and only by this process, until by this process it is decided otherwise.

[1] *Op. cit.* (above, p. 10, note 1), p. 233.

Well, it cannot be denied that art. 88′ is incompatible with art. 88. Art. 88′ removes art. 88, just because logically it is incompatible with it. Therefore, if the validity of art. 88′ is to be derived from that of art. 88, we still have an inference in which the conclusion (the validity of art. 88′) contradicts one of its premisses (the validity of art. 88).

14

Having shown that all the attempts to evade or solve the constitutional puzzle described above in sections 1 and 11 have failed, I shall now venture to present my own solution.

Any attempt at a solution must stand by the principle that from the validity of a norm it is impossible to derive the validity of any norm in conflict with *N*. Therefore, the basic norm of a legal system must be unchangeable in any legal procedure. If the basic norm of a system is in fact changed, this change cannot be derived from any rule of competence (amendment) within the system. This view is not remarkable, it accords well with the theory of deductive systems. The chains of reasons and proofs must stop somewhere ; there must be some foundations (axioms) which are the ultimate basis of all deductions and therefore not themselves demonstrable in the system.

If it is accepted that from a norm of competence, no norm contrary to it can be derived, it follows that the idea of *transference of competence* in virtue of this competence is itself unsound. Even if the basic norm gives the highest authority unlimited competence, this latter nevertheless cannot include power to transfer its power from itself to another authority ; or, generally, power to delimit in any way its own competence. If this is not understood, we land in the well-known paradoxes of omnipotence : Is God able to create a stone so heavy that he is not able to raise it ? [1]

The following attempt at a solution is based on the idea that art. 88—that is, generally speaking, the norm that constitutes the highest authority, in this case the constituent authority—is not the basic norm of the system. Although there is no higher

[1] Ilmar Tammelo and K. Jaakko Hintikka have pointed out the similar antinomy involved in the idea of parliamentary omnipotence. They write : " If Parliament can always pass any law whatsoever, then Parliament can and cannot pass a law limiting its law-making competence. It *can* do this because it can pass any law at anytime. It *cannot* do this, because doing this means that Parliament cannot pass any law at any time." " The Antinomy of Parliamentary Sovereignty ", *Archiv für Rechts- und Sozialphilosophie*, 1958, p. 495.

norm instituting a procedure of amendment (for this would constitute a higher authority) there is still a higher norm investing art. 88 with conditional validity, and this norm is the basic, unamendable norm of the system. This attempt at a solution further evades the notion of a self-destroying transference of competence. In its place appears the notion of *delegation of competence*, that is, a derived competence which does not destroy but functions inside the competence from which it is derived.

To clarify these ideas, let us begin by considering some simpler cases.

Let us suppose

N_0 : Obey your parents !

to be the basic norm for the children in a certain family. As we know, a transference by a parental decision of their power to either another authority replacing the parents or to the children themselves (emancipation) is a logical absurdity. But nothing prevents the parents in virtue of N_0 and in unbroken possession of the highest power from *delegating* power to others. So, *e.g.* if the parents issue the norm

N_1 : During our absence you shall obey Miss A.

This is a time-limited delegation. Its validity is derived from N_0, but N_1 in no way cancels or restricts the basic norm. The parents may at any time, *e.g.* by telephoning to the children, revoke the delegation and restore their actual administration of power.

Now let us imagine some other cases of delegation, *e.g.*

N_2 : During our absence you shall obey A ; if A leaves, before we are back, you shall obey B.

This delegation is conditional as well as time-limited. It is easy to imagine variations in determination of the circumstances that put an end to A's delegated authority as well as in the circumstances that point out B as A's successor, *e.g.* :

N_3 : During our absence you shall obey A ; if A gets intoxicated (or : gets sick ; or : wins in the lottery ; or : marries ; or : uses abusive language ; or : mentions a number higher than 100 ; etc., etc.) you shall obey A's sister (or : the neighbour to the left ; or : a person designated by the drawing of lots ; or : a person appointed by the City Council, etc., etc.).

There seems to be no reason why the fact that puts an end to A's

authority and calls for B as A's successor could not be a declaration to this effect from the part of A. We would then get :

N_4 : In our absence you shall obey A, until he himself points out B as his successor ; from that moment you shall obey B.

N_4 involves no reflexivity and its derivation from N_0 no contradiction. The inference is as follows :

N_0 : Obey your parents !
Fact 1 : Your parents have commanded N_4 ;
Conclusion 1 ; You shall follow N_4 which demands that under certain conditions you shall obey B ;
Fact 2 : The said conditions are now realized ;
Conclusion 2 : You now have to obey B.

Because we are concerned with delegation, not transference, of competence, conclusion 2 is not in conflict with N_0. This is and remains the basic norm of the system. The duty to obey B does not cancel the duty to obey the parents ; it is valid merely inside the scope of parental authority ; at any time the parents may revoke B's authority.

It makes no difference that the chain of delegated authorities has more members ; nor that it is made indefinite (although not infinite), for example as follows :

N_5 : During our absence obey A, until he himself points out B as his successor ; then you shall obey B, until he himself points out a successor, and so on indefinitely.

N_5 involves no infinite regress that denudes it of meaning. At any time, *e.g.* when authority D is in power, we know exactly what fact in virtue of N_5 will terminate the rule of D and institute E in power. It is true that this fact cannot be determined beforehand, because each succession gives N_5 new meaning. From a logical point of view there is nothing objectionable in such a construction. It can be compared with a law that makes it a crime to mention a number higher than a number mentioned by anyone before. Each committed crime then defines a new crime.[1]

In the supposed system N_5 is valid, because it is derived from N_0, the supposed basic norm accepted by the children. We could, of course, imagine another system in which N_5 itself was accepted unconditionally, that is, as the basic norm.

Now, my idea is that the constitutional puzzle is solved if we assume the existence of a basic norm of this pattern as the

[1] This is shown convincingly by Hart, *op. cit.* pp. 309-310.

ultimate foundation of the validity of a legal order containin rules for the amendment of the constitution—such as art. 88 the Danish Constitution. The basic norm would be of this style :

> N_0: Obey the authority instituted by art. 88, until this authority itself points out a successor; then obey this authority, until it itself points out a successor; and so on indefinitely.

It does not matter that this norm refers to an authority created by art. 88, whereas previously I have usually spoken of the article in terms of rules of amendment. The rules of amendment define a law enacting process and that is the same as setting up an authority ; what is created by this process (by this authority) has force as valid constitutional law.

If N_0 is accepted as the basic norm of the system we are able to understand an amendment of art. 88, according to the procedure prescribed by this article itself, as a legal enactment that is valid not in virtue of art. 88 itself, but in virtue of N_0, the basic norm. N_0 itself remains the legally unchangeable basis of the system. On this hypothesis our interpretation of the amendment rules involves no reflexivity and the derivation of art. 88′ from art. 88 no contradiction. It enables us to express without logical absurdities and contradictions the ideas which actually govern the behaviour of people ; and this ability is at the same time the fact which legitimates the assertion that this norm actually *is* the basic norm of the Danish legal system.

University of Copenhagen

METAPHILOSOPHY
Vol. 3, No. 1, January 1972

SELF-REFERENTIAL INCONSISTENCY, INEVITABLE FALSITY AND METAPHYSICAL ARGUMENTATION

JOSEPH M. BOYLE JR.

When one does metaphysics, what exactly is one doing?—this question has come in for much discussion in recent years. In particular, there have been questions about the meaning of metaphysical statements and the validity of metaphysical arguments. This questioning leads to a further perplexity—concerning what it means for metaphysical arguments to be valid.

This essay is an attempt to begin the resolution of this perplexity. An investigation of a kind of argumentation that many philosophers regard as uniquely metaphysical seems a likely starting point for such a resolution. Therefore, in this paper I will analyze a form of philosophical argument widely used by metaphysicians, with a view towards discovering the sense—if any—in which it may be regarded as "valid".[1] This analysis will throw some light on the more general issues concerning the meaning of metaphysical statements and the nature of metaphysical inquiry.

The argument-form that I will examine is that in which a position is criticized as undercutting itself. A position attacked by such an argument allegedly denies or cannot account for some condition that is required for it to make sense or be true. This argument-form has been called the argument from self-referential consistency. It seeks to show that a position or theory which refers to itself, that is to say, includes itself in its subject matter, cannot account for itself. Such a theory can be called self-referentially inconsistent or self-refuting. More colloquially, a philosopher who holds such a position can be said to have cut off the limb on which he sits. A well-known example of this kind of argumentation is the contention that

1See Frederic Fitch, "Self-Reference in Philosophy", *Mind*, LV, 1946, pp. 64-73; Henry W. Johnstone, Jr., *Philosophy and Argument*, (State College, Pa., 1958); Jørgen Jørgensen, "Some Reflections on Reflexivity", *Mind*, LXIII, 1953, pp. 289–300; Richard M. Rorty, "The Limits of Reductionism", in *Experience, Existence and the Good*, ed. I. Lieb, (Carbondale, 1961), pp. 100-116; Russell and Whitehead, *Principia Mathematica*, Volume 1, 2nd ed., Cambridge, 1927); Wilbur M. Urban, *Language and Reality*, (London, 1939); *Beyond Realism and Idealism*, (London, 1949); Paul Weiss, "Cosmic Necessities", *Review of Metaphysics*, IV, 1951, pp. 359-375.

universal skepticism must be rejected because it casts upon itself the same doubts that it casts on other kinds of knowledge.

Aristotle's refutation of the statement that the same thing can both be and not be, is another example of this kind of argument. In Book IV (Gamma) of his *Metaphysics* Aristotle argues that a philosopher who holds this kind of position cannot say anything without thereby showing the impossibility of his view. He cannot say anything because on his grounds words can have no significance. On the assumption that the same thing can both be and not be, whenever a given word is used, its contradictory can be used with equal propriety. Thus whatever we designate as A may rightly be designated as not-A. In short, this position implies that words can have no definite meanings—or even the minimum condition for definite meaning. But if the philosopher holding this view seeks to assert it, he must say something with a meaning that is not wholly indeterminate. In asserting his position he makes use of what he explicitly denies; while disowning reason he listens to it.[2]

A more recent use of this kind of argument is Wilbur M. Urban's refutation of what he calls the naturalist account of knowledge. On this account knowledge is a matter of adjusting to the environment. Urban argues that if the naturalist thesis is taken as an account of *all* knowledge, the naturalist cannot claim that his thesis is true, as in fact he must. The naturalist cuts off the limb on which he sits because by his own theory his thesis cannot have truth value but only survival value. Like every other bit of knowledge, the naturalist thesis is merely the function of the adjustment of the organism to its environment. It has no more significance than any other adjustment.[3]

The inconsistency which is exploited by the form of argument used by Aristotle and Urban may be called "performative", since it lies between a proposition and some aspect of its utterance. To be precise, the inconsistency here is between what a sentence says or expresses and one of the factors operative in its concrete use in discussion or inquiry. The meaning of the statement—i.e., what the sentence is used to express—is incompatible with one of the factors operative in the actual assertion of the statement.[4] The inconsistency which Aristotle's argument

[2]Aristotle, *Metaphysics* 4, 1005B35-1006B35, Chapter 4.
[3]Urban, *Beyond Realism and Idealism,* p. 236.
[4]I shall use the words 'sentence' and 'statement' according to what seems to be established usage. A sentence is a form of words, a grammatical entity that involves no reference to a particular utterance. A statement is an event in which a declarative sentence is uttered. See Jaakko Hintikka *Knowledge and Belief* (Ithaca, 1962), p. 6.

exploits is clearly performative: the thesis he disputes implies the impossibility of expressing a definite meaning, but for it to be used in meaningful discourse it must itself have a definite meaning. Likewise with Urban's argument: the rub in the naturalist account of knowledge is between the general view of knowledge it articulates and the epistemic claim that is operative in its own utterance.

The peculiar character of performative inconsistency can be revealed by comparing it with other kinds of inconsistency. Performative inconsistencies are not purely formal inconsistencies like those of elementary logic. Nor are they merely empirical inconsistencies like the discrepancies that can arise between statements and the facts they purportedly describe. Finally, these arguments neither point out nor depend upon semantical meaninglessness. In short these arguments and the inconsistencies they reveal are not simply matters of syntax or semantics, nor are they simply empirical arguments based only on an appeal to contingent facts. Since the inconsistency here is between what a sentence expresses and its actual utterance, this inconsistency might be called a matter of pragmatics. But there are some difficulties with this characterization because of the historical lineage of "pragmatics".

The irreducibility of performative inconsistency to the syntactical contradictions between elements of a formal system is easily shown. The inconsistency of performatively inconsistent statements is removed if their scope is limited so that the self-referential instance does not arise. For example, Urban's refutation of naturalism works only if the naturalist thesis is taken as an account of *all* knowledge. As an account of lesser generality it makes perfectly good sense and may be very useful. For example, such a theory may be used to good purpose by biographers and historians. As Henry W. Johnstone, Jr. has pointed out, the naturalist thesis claims in effect that people are not responsible for what they say. If the range of the class of people in question is delimited so as not to include the speaker, the thesis is perfectly sensible; it ceases to be sensible and becomes instead paradoxical only when the author claims—whether implicitly or explicitly—that he himself is not responsible for his thesis about people's responsibility.[5]

Unlike purely formal inconsistencies, therefore, the sentences that are involved in performative inconsistencies are themselves coherent and sensible. Jaakko Hintikka has brought out this

[5]Johnstone, *Philosophy and Argument*, p. 71.

difference between performative and formal inconsistency in a striking manner. The performative inconsistencies he deals with are singular statements and not general theses that have self-referential instances; but this difference does not make Hintikka's analysis irrelevant to my interest because these general theses can be performatively inconsistent only in the self-referential instance. Hintikka develops an interpretation of Descartes' *Cogito* argument as a case of performative—and more specifically, of existential—inconsistency. According to Hintikka, Descartes' argument comes down to the fact that he cannot consistently deny his own existence. It is not that there is anything wrong with the *sentence* 'Descartes does not exist' any more than there is anything wrong with 'Homer does not exist'. These sentences become peculiar and paradoxical only when Descartes or Homer respectively tries to utter them. These sentences are perfectly correct as sentences. They become absurd when certain people try to state them, since if such a statement were true, it would imply that its speaker did not exist.[6]

This same point—that performative inconsistency is not reducible to formal inconsistency—is shown by Hintikka in his discussion of another kind of performative inconsistency. He claims that Moore's paradox of saying and disbelieving is performatively indefensible. The sentence 'P but I do not believe P' though logically very peculiar is not logically inconsistent, that is, self-contradictory. Unlike contradictions, a mere change of person makes this into a perfectly normal sentence. 'P and not P' remains a contradiction even when the person of the subject of P is changed. On the other hand, 'P but *he* does not believe P' and most sentences of the form 'P but A does not believe P' are perfectly correct and natural. The inconsistency arises only when the person indicated by A tries to assert this sentence. That is to say, this inconsistency does not depend solely upon the form of words uttered but also on the speaker of the statement. It arises because a speaker must be able to believe what he asserts and this is just what the statement of Moore's paradox denies.[7]

These considerations show that performative inconsistencies are not reducible to the contradictions of elementary logic—they

[6]Jaakko Hintikka, *"Cogito ergo Sum*: Inference or Performance", in *Meta-Meditations: Studies in Descartes*, ed. Sesonske and Fleming, (Belmont, Calif., 1965), pp. 56-60.

[7]Hintikka, *Knowledge and Belief*, pp. 64-78.

are not purely formal inconsistencies. Further considerations show that performative inconsistencies are not examples of semantical nonsense. Semantical nonsense arises when it is impossible to indicate what it is that one is talking about. The semantical paradoxes arise from this impossibility. The statement of the "Liar paradox", for example, predicates falsity of itself but since this statement is nothing other than the predication of falsity, we never find out what statement we are claiming to be false. In this case there is a real sense in which we do not know what we are talking about. This difficulty does not arise in cases of performative inconsistency. Although performatively inconsistent statements refer to themselves, they do not do so in the fashion of the semantical paradoxes. These latter refer to themselves precisely *as referring;* the object referred to is itself a referring phrase and thus to determine the meaning of the original referring phrase it is necessary to determine what its object refers to. Since it too refers to itself, it is clear that the search for a meaning goes on—or around in a circle—indefinitely.[8]

Alf Ross has devised a technique that shows clearly that what he calls "genuinely" self-referential statements, that is, those that refer to themselves inasmuch as they are referring phrases, involve an endless search for their meaning or referent. He calls this a "method of transcription" and it operates on the assumption that the meaning of a word is incomplete until its referent is indicated. This method consists of substituting for the original referring phrase some indication of its referent. For example, 'This man is wise' is transcribed as 'This man—that is, the man John Brown—is wise'. The application of this method to "genuinely" self-referential statements reveals their meaninglessness. The statement 'This sentence is false'—when used to refer to itself—contains a referring phrase in its transcription. 'This sentence—that is, the sentence "This sentence is false",—is false' does not tell *which* sentence is false; nor, it is clear, does any further transcription.[9]

Ross goes on to point out, however, that there are other ways in which a statement may refer to itself. In the concrete, statements are intelligent activities, that is, uses of language for certain purposes according to certain rules. These rules

[8]See Jørgensen, "Some Reflexions on Reflexivity", pp. 289-300: Alf Ross, "On Self-Reference and a Puzzle in Constitutional Law", *Mind,* LXXIII, 1968, pp. 8-13.
[9]Ross, "On Self-Reference" p. 9.

include the rules of grammar and formal logic.[10] Furthermore the utterances of statements are psycho-physical events such as sounds and scribbles. It is clear that when a statement refers to itself in any of these ways, there is no problem in determining what it is about. For example, a statement of grammar can refer to itself as an instance of grammar. There is no difficulty here in determining what it is that one is talking about; such a sentence does not refer to itself *qua* referring but as an instance of grammar.

Thus, the statement 'I remember nothing at all' is paradoxical but not because it has no clear referent. It is paradoxical because the utterance of this proposition—or any other for that matter—requires that its user remember at least how to use language correctly. This general statement refers to itself—that is to say, there are instances of this general statement that refer to aspects of its utterance. These aspects can be readily indicated. They are not elusive like the referents in semantical paradoxes.

Therefore performative inconsistency is not a kind of semantical nonsense. As D. J. O'Connor says: it is one thing for a proposition to predicate falsity of itself and quite another for it to be falsified by its own structure.[11]

This recognition of the irreducibility of performative inconsistency either to the contradictions of logical theory or to the nonsense of the semantical paradoxes undercuts the most common objection to the use of arguments that exploit performative inconsistencies. The objection—formulated by Bertrand Russell and others—assumes that all self-reference is vicious and consequently that all statements that refer to themselves are meaningless.[12] In other words, if all self-referential statements are meaningless, it is impossible to get any philosophically interesting results from the refutation of such statements. On these grounds, then, the only legitimate approach to these statements is to indicate their meaninglessness. Thus, arguments that seek to exploit self-referential inconsistencies are invalid because they try to establish something by arguing with nonsense.

But, as we have seen, not all self-reference is vicious; there are different kinds of self-referential statements. There is no *a*

10See *ibid.*, pp. 10-12; D. J. O'Connor, "Pragmatic Paradoxes", *Mind*, LVII, 1948, pp. 358-359; L. J. Cohen, "Mr. O'Connor's Pragmatic Paradoxes", *Mind*, LIX, 1950, pp. 85-87; C. K. Grant, "Pragmatic Implication", *Philosophy*, XXXIII, 1958, pp. 309-310, 320.

11O'Connor, "Pragmatic Paradoxes and Fugitive Propositions", *Mind*, LX, 1951, p. 536.

12Bertrand Russell and A. N. Whitehead, *Principia Mathematica*, p. 38.

priori reason to assume that those statements whose self-reference is performative are meaningless or nonsensical. Furthermore, there is no reason to suspect that performative inconsistency indicates nonsense. As I pointed out above, the irrationality or absurdity of performatively inconsistent statements is neither syntactical incoherence or semantical lack of reference.

These differences among the kinds of self-reference also suggest that there is no need and no warrant for invoking a theory of types or some other stratification of language to avoid all forms of self-reference. Moreover, the theory of types cannot be regarded as a general prohibition of self-reference without getting involved in the very difficulties it is designed to avoid. As a general prohibition of self-reference it becomes itself an unrestricted thesis about all language and thus meaningless on its own grounds.[13]

This undercutting of a common criticism of self-referential arguments, however, is not enough to establish that these arguments are in fact valid. To establish this validity, the root of the inconsistency must be revealed. Some hints about this basis have emerged in the preceding discussion. We have seen that a performative inconsistency is inconsistency between what a sentence expresses and some aspect of its actual assertion. The question, then, is how what a sentence expresses can be inconsistent with an utterance. I submit that the "inconsistency"—if this is not a completely misleading word in this context—is a case of the discrepancy that obtains between a statement and a state of affairs that falsifies it. It is an example of the inconsistency that arises between a fact and a statement that does not adequately describe it. In other words, a statement is performatively inconsistent when it is *falsified* by some aspect of its utterance.

Talking about a statement being *falsified* by its own utterance may appear, indeed, to be peculiar, but it is perfectly legitimate. A statement involves much more than the sentence and what the sentence is being used to express. A statement is a psychophysical event of some sort. It is also a use of language that makes some sort of claim or proposal. There is nothing to prevent a statement from being used to refer to the event, use or whatever else makes up its own utterance. As in the case of other descriptive statements, the possibility that such a self-referring

[13]Fitch, "Self-Reference in Philosophy", pp. 64-73; Paul Weiss, "The Theory of Types", *Mind*, XXXVII, 1928, pp. 338-348.

statement articulates may or may not obtain. In short, certain statements may be falsified by their own utterance.[14]

An examination of some performatively inconsistent statements shows that it is plausible to regard them as falsified. Urban, for example, is attacking naturalism by pointing out a counter-instance to the naturalist account of knowledge. The fact that the naturalist in uttering his thesis is making some sort of epistemic claim provides an example of knowledge that is at odds with his general account of knowledge. Urban is citing a *fact*—provided in the utterance of the statement—that falsifies the general thesis.

Likewise in the case of 'I remember nothing at all': any use of language requires some memory; the uttering of this statement, therefore, requires some memory. Thus, it provides a counter-example to the general statement.

A similar analysis can be made concerning the other performative inconsistencies that I have considered—even those that do not appear to be examples of falsification. The performative inconsistencies discussed by Hintikka, for example, may not appear to be cases of falsification. Nevertheless, 'I do not exist' can be regarded as falsified. The very utterance of this statement provides a state of affairs that is excluded by the statement: non-existent speakers cannot utter statements. Moore's paradox can be treated in a similar way. Assertion is a certain way of uttering a sentence; it presupposes at least that we can believe what we assert. This possibility of belief is precisely what 'P but I do not believe P' excludes. Therefore, if someone tries to utter this statement as an assertion, the utterance provides the state of affairs that is excluded in the statement.

My interpretation of the source of performative inconsistency appears, however, to generate as many difficulties as it solves. Most of the philosophers who have dealt with this kind of inconsistency have refused to accept the conclusion that it is really a kind of falsification. Johnstone, for example, claims that this kind of argument exploits an inconsistency between what a philosopher says and his fundamental commitments.[15] Hintikka says only that performatively inconsistent statements are "pointless". They disqualify themselves for logical or quasi-logical reasons.[16]

The reason for these refusals is not far to seek. Falsification

[14]See L. J. Cohen, "Mr. O'Connor's 'Pragmatic Paradoxes'," pp. 85-86.
[15]Johnstone, *Philosophy and Argument*, pp. 50-51, 74-75.
[16]Hintikka, "*Cogito ergo Sum*: Performance or Inference", pp. 58-59; Knowledge *and Belief*, pp. 67, 72.

is a factual-empirical notion. It occurs when a situation is not adequately described by a statement that purportedly describes it. This situation is some sort of a fact or state of affairs outside of the statement. The difficulty, of course, is that factual statements—and the statements of their falsifications—are merely contingent. A statement which is both semantically and syntactically meaningful but which is falsified only *happens* not to be true. It articulates a genuine possibility but one that does not now happen to obtain. In other words, the truth or falsity of descriptions is a contingent matter. The only way to find out whether or not a description is true is to look at the world and find out.

The significant characteristic of performatively inconsistent statements, however, is that they are not contingently false. Even though they are meaningful and thus articulate genuine possibilities, they cannot possibly be true. As D. J. O'Connor points out, this is precisely the peculiarity of "pragmatic paradoxes" which are cases of performative inconsistency. They cannot be true yet they are not rejected for any purely logical reason. According to O'Connor the falsity of 'I am not speaking now' is of a "totally different character" from that of 'Churchill is not speaking now' when this is uttered during a Churchill broadcast.[17] The first cannot conceivably be true as the latter clearly can.

This allegedly "totally different character" of performative inconsistencies, on the one hand, from contingent falsities on the other, may explain why philosophers like Hintikka and Johnstone do not consider that these inconsistencies may be a kind of factual falsification.

Certain peculiarities of the utterances of performatively inconsistent statements, however, indicate that they may serve to falsify the statements of which they are a part in such a way that they show them to be necessarily false. These peculiarities revolve around the fact that the utterance of a sentence accompanies the sentence in its concrete uses in discourse and inquiry. The facts that falsify a performatively inconsistent statement are given along with it as it is uttered. Consequently, there is a sense in which these facts are inescapable—they cannot be avoided whenever the statement is to have a real use. Thus, the necessary character of performative inconsistencies arises because the falsifying factors are given as part of the performatively inconsistent statement in its concrete

17O'Connor, "Pragmatic Paradoxes and Fugitive Propositions", p. 536.

occurrence, and not because the inconsistency is not factual.

This is not to say that all aspects of utterances are equally necessary and unavoidable. Some can be easily removed. For example, 'I cannot properly construct an English sentence' is performatively inconsistent. If stated in French, however, it is perfectly sensible. Other features of the utterance are harder to avoid; a statement need not be in English or French but it must be stated in *some* language. Similarly, a statement need not be spoken or written but it must be expressed by some sort of symbol. Clearly, it is these more pervasive characteristics of utterances which the philosophically interesting rejections of performative inconsistency exploit. If what the sentence of a certain performative inconsistency is used to express can be stated in a way that avoids the inconsistency, then this performative inconsistency is not philosophically interesting. Its falsity is dependent upon an idiosyncrasy of a certain way of uttering the sentence and can be removed by uttering it another way. In other words, there is no philosophical importance in those performative inconsistencies whose inconsistency turns on an aspect of the utterance which is properly designated by a proper name or by what functions as such.

On the other hand, those performative inconsistencies which cannot be stated in a way that avoids the inconsistency are most interesting because what they state is inevitably falsified. There is no concrete context from which the falsifying factor is absent. These performatively inconsistent statements are inevitably false. Therefore, the basic reason for refusing to admit that performative inconsistency is a kind of factual falsification is not an adequate reason. The *necessity* for rejecting performatively inconsistent statements is not at odds with the factual basis of this rejection because the falsifying facts are present whenever the statement is uttered.

The adequacy of my account of performative inconsistency is confirmed by the difficulties inherent in the other accounts of it. The "pointlessness" which Hintikka cites as the basis for performative inconsistency is not specific. Any kind of contradiction or inconsistency renders a statement pointless. For example, if the point or purpose of a statement is taken to be communication, it is as well thwarted by the statement of syntactic inconsistency or of semantic nonsense as by the statement of a performative inconsistency. Of course, the question about the basis of performative inconsistency concerns

what is specific to it—not what it has in common with other kinds of inconsistency.

Johnstone's contention that performative inconsistency lies between a statement and a philosopher's commitments also involves difficulties. It is clear that there are performatively inconsistent statements whose inconsistencies do not turn on the commitments implicit in their utterance, but on other aspects of the utterance. For example, Urban's refutation of naturalism, which Johnstone explicitly considers, does not exploit the naturalist's commitments but the fact that he makes a truth claim—or some other claim of cognitive adequacy. Certainly commitments are operative here but they are not the naturalist's specific commitments to naturalism, nor are they basic in generating the inconsistency. The naturalist's commitment to give a completely general account of all knowledge gives rise to the self-referential instance whose statement is performatively inconsistent. But this commitment to give a general account is not a uniquely naturalist commitment nor does it provide the basis of the inconsistency. It provides the instance of the generalization that is performatively inconsistent—not the basis of the inconsistency of that instance.

At this point another sort of objection may arise precisely because of the similarities between performative inconsistency and factual falsification. If I have succeeded in showing that performative inconsistency is not merely a matter of self-contradiction or semantical nonsense, then I may be faced by the challenge that it is really only a factual mistake.

First of all, it may appear that like other facts those that falsify performatively inconsistent statements are susceptible of a wide range of possible interpretations. Thus susceptible of many interpretations, the factors that could falsify such statements might be interpreted so as to avoid the inconsistency. This possibility seems to neutralize the unavoidable character of these falsifying factors. A fact that can be interpreted away—however unavoidable it may be—is not intractable or obstinate; such a fact could hardly be the basis for a compelling and conclusive philosophical argument.

The objection, however, fails because the interpretation of the utterance must itself be uttered. It too can be performatively inconsistent and it surely will be if it is an attempt to interpret away the embarrassing aspect of the utterance.

In other words, the new interpretation of the utterance can be a merely verbal and *ad hoc* device for avoiding the con-

sequence of a valid argument. This becomes clear if the same issues which forced the reinterpretation arise again in the utterance of the statement of the reinterpretation.[18]

Thus it is clear that interpretation can be useful here only to show that what was conceived as a performative inconsistency is not really such. If used as an evasive technique it fails; the inconsistency comes up again—in the utterance of the interpretation. This is not to say, of course, that a philosopher accused of performative inconsistency cannot interpret his utterance so as to avoid its generating the inconsistency. But this forces him to change his original statement as well if it is not to be inconsistent with the statement of his interpretation. These two statements will be inconsistent because the latter describes what falsifies the former. This kind of interpretation does not evade or obviate a self-referential criticism. Rather, it alters a position to avoid such a criticism. Thus it presupposes the validity of the criticism.

But this discussion seems to provide no criterion whereby the philosophers in such a dispute about the interpretation of the utterance can systematically avoid begging the question. For example, who is to determine whether the interpretation is performatively inconsistent? The interpreter or his opponent? This ever-present possibility of begging the question is another side of the issue raised by the possibility of interpreting facts in a variety of ways. Like it, this danger is a problem in any empirical argument. The problem is that since aspects of the utterance falsify performatively inconsistent statements, these must be recognized and described at least in some way. But recognition and description appear to involve interpretation.

The objection, however, does not show that self-referential arguments are inevitably question-begging. In the first place, the proper use of this kind of argumentation does not demand that the philosopher who is attacked by it explicitly recognize or understand the factor that generates the performative inconsistency. This follows from the fact that the factors involved in the utterance of a statement are operative in the performance of uttering the statement even if they are ignored by the speaker. These factors are implicit in what he is *doing* even if he does not explicitly refer to them. It follows from this that

[18]See Richard Rorty, "The Limits of Reductionism", pp. 104-110. The context of Rorty's discussion is different from mine. He considers the case where the charge of self-referential inconsistency is accepted and parried by a distinction of level, that is, a distinction that includes the original distinction as one of its divisions. This allows the philosopher to deal with his subject without having to consider his analysis. *Mutatis mutandis* my point is the same as Rorty's.

it is not question-begging to use this kind of argument against a philosopher who does not recognize the factor that generates the performative inconsistency of his position.

Secondly, the use of this argument does not depend upon an interpretation of the falsifying factor by the philosopher who uses it for polemical purposes. It is enough for him to recognize that the statement he is attacking is inconsistent with some aspect of its utterance. This does not require a definition or account of this aspect of the utterance. Conversely, the successful use of this argument does not justify a philosopher's interpretation of the factor that generates the performative inconsistency.

The fact that very often performative inconsistencies initially present themselves as *paradoxes* further confirms my contention that an interpretation of the falsifying factor is not necessary to expose a performative inconsistency. With paradoxes we discern that something is wrong even before we can identify with precision the source of the difficulty.

Another way of showing that the uses of this argument need not be question-begging is by examining some examples. Urban's refutation of naturalism is a good example because it can appear to be question-begging. First of all, it is clear that Urban's argument does not depend upon what the naturalist is willing to admit. In fact, Urban is contending that the naturalist is doing something, that is, making epistemic claims, which is inexplicable on his theoretical grounds. In other words, the naturalist is doing something he cannot admit. Furthermore, Urban's argument does not depend upon his own account of what constitutes a truth claim. He seeks to show that, whatever a truth claim is, it is not simply adjusting to the environment. Should the naturalist admit that his claim is merely this, he would surrender all claim to plausibility. But he *is* claiming plausibility: he argues and presents evidence.

The "paradox" of the naturalist position can be revealed without any explicit mention of epistemic adequacy and its alleged prerequisites such as responsibility. This paradox can be shown by the possibility that someone should explain the theory in its own terms. It follows, of course, that this explanation of the theory in its own terms can also be explained in terms of the theory. This explaining of the explainings of naturalism can go on indefinitely, and at some point the paradox involved in this possible progression emerges. It emerges because at each step the explanation disqualifies those before it from being cognitively adequate—at least according to most understandings

of cognitive adequacy—but the paradox can emerge without any clear understanding of this reason. If the naturalist is willing to admit that cognitive adequacy is just a matter of environmental adjustment, he must be careful to make no claims that cannot be construed as environmental adjustments. For example, how can a philosopher who understands knowledge in this way expect that others be obliged to believe him? Their evaluative response—itself an environmental adjustment—could be quite different from the naturalist's and remain an effective adjustment. In other words, the philosopher who understands epistemic adequacy in this way cannot exclude as inadequate alternative or opposed positions on any subject on which he claims epistemic adequacy.

In short, no single definition of 'truth claim' is needed to see the absurdity or paradox of the naturalism which Urban attacks. On the contrary, the oddness of a certain account of 'truth claim' is revealed in this paradox. The naturalist either makes a claim that his theory renders impossible or so weakens the notion of epistemic claim that he precludes his own credibility.

As I have already noted, Urban's argument is one that can appear to be question-begging; other arguments that exploit performative inconsistencies are not even apparently question-begging. The performative inconsistencies discussed by Hintikka are of this sort. The refutation of 'I do not exist' depends upon the paradox inherent in the idea of a nonexistent person performing an activity. Various interpretations of the activity of uttering a statement or of the person who utters a statement do not affect the paradox of this statement. The rejection of 'P but I do not believe P' is even more clearly not question-begging. It is not that one has an account of assertion that implies that the speaker be able to believe what he asserts; rather, the very activity of asserting Moore's statement conflicts with what the statement asserts. Thus, we may conclude that assertion requires that the speaker be able to believe what he asserts not on the basis of our theory of assertion but because of the paradox that arises when we try to exclude this possibility.

Among the uses of this type of argument that avoid the appearance of begging the question are those that attack various dualisms. For example, the dichotomy of everything into reality and appearance after the fashion of Parmenides' monism is performatively inconsistent. This dichotomy leaves no room for making the kind of claim that one makes in stating the

dichotomy. The position cannot be articulated from the favored side of the distinction—from within reality—because true reality is absolutely unified, and the statement of the theory involves multiplicity—of words, of concepts, or of language and reality.[19] Nor can the thesis be articulated from the appearance side of the dichotomy: from this point of view the distinction between reality and appearance is merely apparent. In short, the philosopher like Parmenides has left himself no possible standpoint from which he can articulate his thesis; he has no position from which to discern the basis for his claim. This argument is clearly not question-begging: the meanings of 'reality' and 'appearance' are given by the philosopher under attack. The possible interpretations of utterances do not affect the factor that is operative in generating the performative inconsistency here—namely, the fact that the utterance of a proposition involves some sort of multiplicity.

Another dichotomy that can be attacked as performatively inconsistent is the division of all meaningful statements into empirical descriptions and tautologies on the basis of the verifiability principle. It is clear that the statement of this principle can fall into neither of these categories; it is certainly not a description or empirical generalization, and if it is taken to be a tautology it cannot do the job it was meant for. If it is regarded as a definition or prescription, and as such, a kind of tautology, it cannot be used for polemical purposes. There is no need for other philosophers to accept this prescription or definition. Thus, the statement of the verifiability principle—whatever sort of statement it is ultimately found to be—is not a tautology or description; it is not a meaningful statement according to its own criteria. But it is clear that in some sense of the word it must be meaningful; the verifiability principle must make sense to its proponents and their opponents. Otherwise, neither side in the dispute is really saying anything.

To sum up: these examples confirm my contention that arguments that exploit performative inconsistencies need not be question-begging. This conclusion further suggests that these arguments can be regarded as valid. Once the basis of the inconsistency on which these arguments turn is revealed, the problem of begging the question is revealed as the basic difficulty. As we have seen, other objections assume that performative inconsistency is a purely formal or semantical matter, or that the self-reference involved is of a logically vicious type.

[19]See Paul Weiss, "Cosmic Necessities", p. 362.

These objections are undercut once the *factual* character of performative inconsistency is understood. Moreover, this realization of the factual character of the inconsistency establishes the argument as having a legitimate basis. The discrepancy between statements and facts that falsify them is recognized as a legitimate procedure for rejecting the statements. But the factual character of the argument introduces the possibility of interpretation and the danger of question-begging. Since the argument need not be question-begging we can conclude that it is valid.

At this point, however, some question may arise as to why I regard this argument-form as peculiarly metaphysical. The examples we have discussed suggest that it can be used in all areas of philosophy. Why characterize it by its use in one of the most disputed philosophical areas?

I call this type of argument peculiarly metaphysical, first, because it can terminate in statements that are certain, and secondly, because it can be used to make statements about the whole of reality or about "being as such". These characteristics define at least some of the criteria for metaphysical statements. A word of explanation is required, however, especially about the notion of "certainty". As I am using the term, it is not a psychological category but an epistemic one. It does not refer to a state of mind but to the kind of knowledge one has of a statement that is known to be irrevisably true. A certainty, therefore, is a statement that is known to be simply and definitively true. Thus it differs from the way we know descriptive statements since these are at least in some measure contingent upon one's point of view. It differs from the way we know hypotheses and proposals since their adequacy is never a matter of simple truth. It differs from the way we know logical or mathematical "truths" since these are a matter of validity and not truth in the semantical sense. Logical truths say nothing about the way things are. Furthermore, the certainty we achieve here clearly is not an unreal objective of an illusory desire. It is a function of a philosopher's perfectly legitimate concern to achieve a thoroughly critical standpoint. This requires that such a philosopher seek a position that he can know to be independent of his assumptions and prejudices.

The rejections of performative inconsistencies are certainties. They are not hypotheses or descriptions but rather the statements of the falsifications of hypotheses or descriptions. In other words, the terminations of these arguments are denials—negations—and not descriptions or hypotheses. Furthermore,

they are non-revisable denials. The factors that falsify performatively inconsistent statements are unavoidable; the falsification of these statements is *necessary*. A statement convicted of performative inconsistency is simply and definitively false.

Finally, these negations are not based on logical or semantical incoherence. It is not that performatively inconsistent statements are nonsensical and thus do not say anything about the subject-matter; they say something about this subject-matter but are rejected because they are incompatible with some aspect of it. Therefore, unlike the denial of nonsense, negations based on performative inconsistency say something; they state that however the subject in question is to be defined or described, it cannot be in a way whose statement is performatively inconsistent. This is, however, to delimit the subject-matter in some way because it reveals certain coherent descriptions of it that cannot be adequate.

There is still question about how this type of argument can be used to ground statements about the whole of reality. Its role as a falsifying procedure prevents it from being used to articulate or defend any theory or positive understanding of reality. Most important it cannot be used to establish the axioms of some metaphysical system. Such axioms are general theses and the statements of falsifications are not. But it can be used to falsify such axioms if they are performatively inconsistent and this suggests the way in which it can be used to make statements about reality as such.

This kind of argument can show that certain explanations of reality are false if taken as explanations of reality as a whole. Thus it can introduce a kind of complexity into our discussion of being: it can show that whatever being or reality is, it cannot be understood simply by a position whose statement is performatively inconsistent. This is not to give an account or description of being but it is a kind of negative definition or delimitation of the meaning of reality.

This establishment of the validity of an important type of philosophical argument—and the propriety of its use in metaphysical inquiry—has implications for the discussions about the nature of the enterprise of metaphysics. Most important, it implies that metaphysics—at least to the extent that it depends upon this kind of argument—is a rigorous discipline of inquiry which can achieve objective results based on legitimate criteria. In other words, metaphysics is neither a merely poetic expression of a metaphysician's subjectivity nor a matter of

conceptual explication. It is, rather, as many metaphysicians have claimed, knowledge of truth—if only the truth of negations.[20]

[20]This article is based on my dissertation, *The Argument from Self-Referential Consistency: The Current Discussion* (published Ph.D. dissertation, Georgetown University, 1970), which was directed by Germain Grisez. I wish to thank Grisez for his invaluable help in preparing this paper.

AQUINAS COLLEGE

AUTHOR INDEX

Åquist, L. 262
Agassi, J. 121
Aiello, N. 361
Alexander, P. 187, 188
Andersen, P. 463
Anderson, A. R. 168
Aristotle 125, 191, 454, 482
Asher, N. 417, 404, 407
Attardi, G. 368, 384
Austin, J. L. 121, 132, 197, 199, 207
Ayer, A. J. 249, 260

Baker, A. J. 206
Bar-Hillel, Y. 191, 192, 199-201
Barry, G. 229
Bartlett, S. J. 4, 6, 13-15, 17, 92, 260, 261
Beardsley, M. 128
Belnap, N. D., Jr. 168
Bentham, J. 10
Bergson, H. 37, 44
Black, M. 124, 191, 203, 205, 206
Bobrow, D. G. 341, 342
Bochenski, I. M. 125
Bochvar, D. A. 256, 262
Boolos, G. 390, 404, 417
Bouwsma, O. K. 13
Boyle, J. M., Jr. 10, 14
Brachman, R. J. 341, 342
Bradley, F. H. 41, 207
Bridgman, P. W. 13, 253
Burge, T. 368, 384, 417
Buridan, J. 101, 110, 115
Burks and Copi 199

Cantor, G. 32, 33, 222-224, 227
Carnap, R. 75, 77, 78, 119, 121, 191, 249, 260, 265, 266, 273, 274, 282, 283, 467
Carroll, L. 109
Cheshire and Fifoot 193
Chomsky, N. 126
Christensen, N. E. 463, 466, 471, 472, 475, 476
Christiansen, A. 202
Church, A. 6, 75, 76, 97, 116, 167, 168, 262, 266, 282, 283
Chwistek, L. 6, 76
Cohen, L. J. 101, 179, 180, 181, 187-189, 486, 488
Cohen, B. 342
Collingwood, R. G. 3, 16
Collins, A. M. 343, 361
Cousin, D. R. 183
Creary, L. 368, 384
Curry, H. B. 6, 260

Davis, M. 280
De Morgan, A. 22
des Rivieres, J. 397, 418
Descartes, R. 106, 224
Dickinson, L. 226
Donnellan, K. S. 134
Doyle, J. 342, 383
Drapkin, J. 417

Dummett, M. 110, 111
Duncan-Jones, A. 189

Eberle, R. 417
Ebersole, F. B. 187
Einstein, A. 28
Elschlager, B. 368, 384
Emmet, D. 207
Epimenides 25, 27, 30, 35, 39, 103, 464
Eudoxus 9

Fagin, R. 417
Feferman, S. 417
Fink, E. 4
Fitch, F. B. 147, 216, 234, 481, 487
Frege, G. 30, 133, 135, 265, 266, 283, 364, 365, 370, 371, 383

Gallaire, H. 341
Geach, P. T. 99, 101, 105- 107, 110, 191
Gentzen, B. 6
Gettier, E. 417
Gilmore, P. 372, 373, 375, 378, 384, 393, 417
Goad, C. 281
Goddard, L. 132
Gödel, K. 4, 6, 21, 34, 35, 95, 223, 233, 328, 405, 407, 410, 417
Goodman, N. 80
Grant, C. K. 486
Grelling, K. 23, 30, 88, 187
Grisez, G. 10, 14
Grue-Sørensen, K. 466, 468
Gurwitsch, A. 9

Haas, A. 367, 369, 370, 384
Halldén, S. 262
Halpern, J. Y. 358, 359, 361, 417
Hare, R. M. 205
Hart, H. L. A. 466, 471, 475, 479
Hayes, P. J. 280, 283, 383, 384
Hegel, G. W. F. 424
Heintz, J. 168
Hempel, C. 78
Hendrix, G. 368, 384
Henkin, L. 6, 307
Hilbert, D. 447
Hintikka, J. 261, 341, 346, 354, 361, 395, 417, 477, 482-484, 488-490, 494
Hocking, W. E. 10
Hume, D. 10, 249, 260
Husserl, E. 4, 7, 16, 427

Ipsen, D. C. 13
Isaye, G. 16
Israel, D. J. 341, 342

Johnstone, H.W., Jr. 8, 9, 11, 12, 481, 483, 488, 489, 491
Jørgensen, J. 464, 481, 485
Jourdain, P.E.P. 37

Kalmár, L. 6
Kamp, H. 404, 407, 417
Kant, I. 4, 7, 12, 14, 16, 239, 259, 423
Kaplan, D. 276, 282, 283
Kleene, S. C. 6
Kneale, W. and M. 119
Kneale, W. 120, 121
Konolige, K. 346, 361, 366, 369, 383, 404, 411, 417
Kordig, C. R. 10
Kotarbiński, T. 79
Krarup, O. 463
Kreisel, G. 6
Kripke, S. A. 279, 346, 350, 361, 371-373, 375, 377, 378, 384, 393, 398, 404, 417

Lambert, K. 168
Langford, C. H. 210
Lawrence, - 53, 58
Leblanc, H. 341
Levesque, H. J. 341, 342, 350, 361, 397, 417, 418
Lewy, C. 204
Linsky, L. 283
Lipski, W. 341
Liskov, B. 341
Löwenheim, L. 4, 6
Lorenzen, P. 12, 242
Lovejoy, A. O. 10
Łukasiewicz, J. 262

MacIver, A. M. 205
Mackie, J. L. 103, 106-108, 112, 113
Malcolm, N. 191
Martin, R. L. 213
Martin, R. M. 119
McCall, S. 168
McCarthy, J. 283, 341, 361, 280, 281, 352, 366-368, 383, 417
McDermott, D. 342, 383
McTaggart, J. 104
Mendelson, E. 341, 418
Meyer, R. 168
Mill, J. S. 12
Miller, M. 361
Minker, J. 341
Moleski, M. X. 16
Montague, R. 266, 282, 283, 368, 370, 384, 385, 387, 388, 394-396, 398, 400, 404, 406, 407, 410, 415, 418
Moore, G. E. 9, 200, 201, 205, 207, 208
Moore, R. C. 266, 341, 344, 347, 358-360, 362, 368-370, 384, 409, 418, 494
Morgan, C. G. 281-283
Moses, Y. 358, 359, 361, 417
Mostowski, A. 6, 262
Muck, O. 16
Myhill, J. 6

Nagel, E. 34
Nehrlich, G. 168
Newell, A. 285-288, 290, 341
Newman, J. R. 34
Newton, I. 15
Nowell-Smith, P. H. 191, 192

O'Connor, D. J. 175, 176, 179, 181, 191, 486, 489
Opsahl, T. 463

Pap, A. 129, 130, 440, 441
Partee, B. 418
Passmore, J. 10, 12
Péter, R. 50
Peano, G. 79
Peirce C. S. 427
Perlis, D. 383, 417, 418
Pollock, J. 168
Popper, K. R. 466, 468-472
Post, E. 6
Prior, A. N. 115-117, 168
Protagoras 10

Quine, W. V. O. 12, 76, 80, 84, 88, 120, 140, 147, 168, 261, 274-276, 283, 294, 399, 418

Ramsey, F. P. 187, 253
Reinchenbach, H. 174
Reiter, R. 341, 342
Rescher, N. 12, 168
Rich, C. 342
Richman, R. J. 260
Rieger, C. 390, 418
Rorty, R. 10, 481, 492
Ross, A. 485
Rosser, B. 6, 50, 95
Royce, J. 10, 72
Rüstow, A. 121
Russell, B. 15, 21, 26, 28-30, 37, 53, 70, 71, 74, 120, 174, 227, 276, 365, 371, 383, 463, 464, 481, 486
Ryle, G. 73, 275

Schlick, M. 249, 260
Schmidt, P. 260
Segelberg, I. 466
Sellars, W. 124
Shakespeare, W. 211
Shaw-kwei, M. 262
Simi, M. 368, 384
Skolem, Th. 4, 6, 81
Skyrms, B. 168
Smith, B. 388, 418
Smorynski, C. 390, 404, 418
Smullyan, R. 380, 384
Sobociński, B. 102
Sommers, F. 126
Sørensen, M. 463
Sosa, E. 168
Spaulding, E. G. 10
St. Paul 26
Stalnaker, R. 361
Stefik, M. 342
Stich, S. 418
Strawson, P. F. 14, 16, 154, 168, 191, 199, 200, 201, 206, 207, 261
Suber, P. 4, 6

Tammelo, I. 477
Tarski, A. 6, 26, 28, 49, 72, 76-79, 99, 109, 110, 120, 124, 160, 231, 233, 234, 236, 237, 295, 370, 371, 384, 392, 407, 418
Thomason, R. 385, 387, 388, 400, 404, 407, 418
Tollefsen, O. 10, 14
Turing, A. 6

Urban, W. M. 10, 226, 481, 482, 483, 488, 491, 494
Ushenko, A. P. 187
Uspenskij, V.A. 6

Vardi, M. 418
von Neumann, J. 78

Wang, H. 6
Weiss, P. 168, 481, 495
Weyhrauch, R. 368, 384, 414
Weyl, H. 39, 42
Whitehead, A. N. 37, 41, 222, 427, 463, 481, 486
Whorf, B. L. 4
Willis, R. 206
Wilson, C. 210
Winograd, T. 383, 384
Wittgenstein L. 104, 119, 260
Woodruff, P. 168

Yap, C. K. 283

Zeno 22, 23, 28, 29
Zermelo, E. 34, 78
Zilles, S. 341

SUBJECT INDEX

Absolute presuppositions of systematic thought 16
— space 15
— time 15
— truth 15, 46
Abstract expressions 281
— languages 281
— syntax 281
Abstraction 90
Accessibility relation (Kripke) 282
Ad hominem argument 13, 225, 226, 228, 453
— — —, and self-applicability 216
Adequacy 79
Aesthetics 430
Ambiguity, systematic 71
Antinomies 23-25, 28
—, and crisis in evolution of thought 34
—, infinite series of 32
Argumentation, philosophical 443ff
— —, *see also* Philosophical argument
Artificial agents 344
Artificial intelligence 7, 17, 263ff, 368, 379n
— —, environments in 278
— —, study of concepts in 280
Artificial language 72
ASK, definition of 310
Auto-epistemic reasoning 385
Autoepistemic theories 360
Autological adjectives 23
Axiom of equality 298, 327
— of infinity 80
— of reducibility 227
— of specialization 297, 307, 308, 327

Barber's paradox 21, 22, 25, 30
Behaviorism 10
Belief 204, 287, 363ff, 385
—, agent's knowledge of his own 343
— as provability 404
—, flattened, ideal theories of 389
—, formalizations of, inconsistent 387
— in an artificial agent 344
— introspection 343
— set 345
— subsystem 344
— —, *see* Introspective belief subsystem
Beliefs, ideal rational agent's 358
Berry's antinomy 28
— —, *see* Berry's paradox
— paradox 6, 28, 30, 43, 69
Biological homeostasis 6
Bivalence 168
— needed to derive absurdity 165
—, *see also* Law of bivalence
Bracketing (Husserl) 427f
Burali-Forti paradox 6
Buridan's paradox 105

Cantor's paradox 6
— proof 31
— theorem 33

Categorical imperative (Kant) 426
Cause-effect sequence 15
Circular definition 227
Claims, capacity to be meaningful 16
Class, intensional definition of 71
— of classes 44
Classes, distinction between classes and class names 71
—, self-membership of 29
Classical valuation 157
Closed world assumption 303, 333
Closure 304
—, assumption of 306, 309
Co-necessitation, and Tarski's principle 162
Cognitive science 17
Coherence theory of truth 10
Combinatory logic 234n, 235
Common sense 416
— —, defense by Moore 12
— — reasoning 385
— —, introspection in 409
— —, substitution in 399n
Communication of information 128
Compactness theorem 296
Competence 304
—, assumption of 304, 309
—, notion pertaining to knowledge 286ff
Completeness 50, 297
— theorem, for first-order logic 382
Complex predicates, sentences with 146
Comprehension axioms 365
Computational reflexivity 7, 263ff
Computer programs, with general intelligence 280
Concatenation, in diagonalization 84n
Concepts 268, 272, 363ff
—, in artificial intelligence 280
—, as mathematical objects 265
—, individual 265
Conceptual analysis 6
— pathology 17n
— vocabulary 5
Consistency 50, 374
—, pragmatic criterion of 454
—, proof of relative 80
Consistent languages 103
Constants 88
Constitution (legal), as highest authority 458
—, legitimizing succession of legal authorities 480
—, levels of authority 458
—, puzzle of self-amendment 457ff
—, self-amendment 457ff, 480
Constitutional norm of competence 458
— paradox 463
— —, *see also* Constitution (legal), puzzle of self-amendment
— rules of competence 457
Contextual implication 192
Contingent truth 422
— —, convertibility into existential 422
Convergent series 28
Copernican paradox 28

Coreference relation 295
Cosmic formulations, impossibility of 37
— necessities 421, 447
Criteria, logically compelling 248
— of meaning 247, 249, 439
— — —, external 249
— — —, intrinsic 250
— — —, meaningfulness follows from self-application 249
— — —, necessary, not sufficient, condition 257
— — —, non-arbitrary 257
— — —, self-validating 257
—, non-arbitrary 248, 249
Criterion of exemplifiability 454

De Morgan's law 273
—— theorem 237
De-projection, method of 260
Default reasoning 332
Definability, strict 93
Demonstrative reference 134, 135, 139
Denotation 266, 268
Denotational definition 272
— semantics 80
Descartes' *cogito*, as performative 484
— evil genius 13
Descriptions 240
—, denotation of 278
—, referential preconditions of 241
Descriptive identification 135
Determinism 11
Diagonal argument 233
Diagonalization 84n
— lemma 407
— versus normalization 94
Dispassionate survey (Peirce) 428

Egocentric particulars 174, 179
Elementary means of referring 242
— sentences 242
— ways of speaking 242, 244
Empirical self-reference 146
— —, paradoxical 142
Encoding systems 4
Environments, in artificial intelligence 278
Epimenides' paradox 49, 63, 230
— —, *see* Liar paradox; *and* Epimenides *in Author Index*
Epistemological boundaries 4
Epistemology 17
Epithets, epithets of 61
—, philological 55
—, phonetic 55
—, self- 55
Errors, detecting 259
Essential truths, conflict with possibility of themselves 423
— —, types of 424
Esthetics 430
Ethics 430
Excluded middle 168
— —, *see* Law of excluded middle
Existential absurdity, definitory 422
— —, essential 422
Extensionality 269
Extensional logic 46
— propositions 278

Factual commitments 8
Falsidical paradoxes 22, 23, 28
Feedback in argumentation, judo-like strategy 11
—, negative 6
—, positive 6
Feedforward, negative 6
—, positive 6
First-order theories, of concepts and propositions 265
Formal metalanguage 233
Formalized language-system 75
— logistic system 75
— philosophy, and Tarski's theorem 231
Formally undecidable propositions 6
Foundations of mathematics 17
Framework of identification 241
Framework-relativity 17
Frederic paradox 22, 23
Free choice 14
Fugitive propositions 187, 188
Fulfillment, completeness property for introspection 353
Functionalism, in logic 445ff
Functors, modal 99, 104
—, monadic proposition-forming 97
—, non-extensional 99
— of mental attitudes 104
—, truth- 104

General intelligence 280
— propositions, capable of being own arguments 40
— systems theory 17
Generality of reference 122
Gilmore-Kripke schema 393
Gödel correspondences 93
— sentences 83
Gödel's incompleteness theorem 390n
— proof, incompletability of number theory 34
— second incompleteness theorem 406
— theorem 84, 92, 410n
Gödelization, and self-reference 217
Grammatical categories 125
Grelling's antinomy 29
— —, *see* Grelling's paradox
Grelling's paradox 6, 23, 68, 187, 234-236

Henkin sentence 88
Heterological adjectives 23
— paradox 375
Heterological-autological contradiction 42
Heterologicality 53
Hierarchies 38, 60
— of belief subsystems 346
— of formal metalanguages 232
— of language 53
Hierarchy of classes 71
— of limited negations 237
— of theories 389
— of truth predicates 371
Hintikka's refutation of "I do not exist" 494
Human brain 6

Ideal agents 347
— grammar 126
— introspective agent 358
Idealism, views that oppose 11
Identification 251

—, framework of 241
—, metalogical properties of 252
Identifying reference 251
Immunity to revision 10
Implication 153
Incomplete knowledge 292
Incompleteness 87
Inconsistencies, in modal logics 404
Inconsistency 87, 443ff
—, multiple criteria for identifying 455
Inconsistency-engendering predicates 6
Indirect self-reference 122n, 145, 379
— —, paradoxical 141
Inductive inference 364
Inevitable falsity 481ff
Inference, inductive 364
—, temporal 364
Informative discourse, presuppositions of 128
Integers 33
Intensional isomorphism 273
— language 290
— logics 283
— reasoning 270
Intensive propositions 46
Interaction language 286
Interface language 290
— —, extended 299
Internal analysis and criticism 247
— limitations of human understanding 3
Internally inconsistent dogmas 8
Introspection 343
—, always retrospection 73
—, negative 348
—, positive 348
Introspective agent, ideal 358
— belief subsystem 346, 349
— — —, consistency of 346
— — —, decidability of 349
— — —, faithfulness 352
— — —, fulfillment 352, 353
— knowledge, perfect 348
Intuitionism 10
Invariant conditions of discourse 12
Irrevocable falsification 10

Knowing, always a temporal process 73
— that 266
— what 266
Knowledge 385
—, accessing 288
— acquisition 305, 318
—, and memory 287
—, and truth 387
—, augmenting 288
— base, retrofitting existing 355
—, flattened, ideal theories of 389
—, formalizations of inconsistent 387
—, implicit 306
—, incomplete 292
— level 286
— representation 285ff, 313
— — systems 285ff
Kripke accessibility relation 282

Language, extended interface 299
—, intensional 290
—, interaction 286
— interface 290
— of concepts 265
—, self-describing 367
—, self-referential 368, 386
— — —, *see also under* Self-Reference
—, universal 368
—, unqualifiedly substitutive 391
Language-hierarchies 53
Language-levels, hierarchy of 103
Languages, in which self-reference is possible 83
Law of bivalence 159, 255
— of excluded middle 147, 159, 235, 236, 372, 377, 445, 451
— — — —, is "false" 130
— of extensionality 104, 107
— of non-contradiction, as universal criterion of consistency 454
Legal system, basic norm unchangeable in 477f
Levels of language 120, 368
— — —, too restrictive for artificial intelligence 368
Lex posterior principle 476
Liar paradox 119, 217, 469, 472, 485
— —, Eubulidean version 97
— —, generality of reference 143
— —, involving attribution of falsity 121
— —, language-level solution 120
— —, language-level theory not adequate 150
—, strengthened 164
—, weakened 167
Limited negation 236
Limitedly false sentences 130
Limits of possible experience 12
— of possible knowledge 5
Linguistic categories 125
— performances 208
— phenomena, never self-referring 66
— relativity 4
Löb's theorem 388, 406, 408, 410n, 411, 414, 416

Machie's formulae 104
Machine computation 35
Mathematical dissection (Whitehead) 428
Maximal set theorem 316
Meaning criterion 439
— —, *see also* Criterion of meaning
— spectrum 258
Meaning-theory 439
—, self-reference of 439
Meaningfulness condition 116
— of claims 16
Meaningless questions 13
— statements, eliminating 259n
Meaninglessness 26, 37, 64, 115, 248, 262n, 463
—, projective 254
Memory, and knowledge 287
Meta-knowledge 302
Metalanguages, universal, for philosophy 231
Metalinguistic immunity 440
— —, rejection of 441
Metalogic of reference 239ff
Metalogical referential consistency 17
— self-reference 7, 12, 253
— —, constructive use of 15
— —, critical use of 13
Metaphysical argumentation 481ff
— arguments, validity of 481
— self 14, 16
— statements, meaning of 481

Metaphysics 497
—, as knowledge of the truth of negations
Method of transcription 464, 473, 485
— — —, circularity of 473
Methodological principle of validation 45
Mistakes, making 259
Modal functions 272
— logi , specifically designed for artificial intelligence 397n
— —, unqualifiedly substitutive 403
— logics 272, 385
— —, inconsistencies in 404
Model 158n
— theory 266
Modus ponens 154, 160
— tollens 154, 155, 160
Monism (Parmenides), refutation of 494f
Monotonicity in real agents 352
Montague's theorem 412n
Moore's paradox 484, 488
— refutation of "P but I do not believe P"
Motivation, in philosophical argument 449
Multiplicative axiom 80

Names 77
Naming relation 187
Naturalism, Urban's refutation of 482, 488, 491, 493
Necessities, cosmic 421
Necessity, existential 421
—, logical 421
Negation, limited 236
Negative introspection 348
— science (Kant) 259n
Neurological limitations 4
Nominalism 222
Non-rational ideas, people dominated by 462
Non-self-membership 34
Non-translational semantics 75, 77
Nondoxastic formulas 345
Norm 88n
— function 83, 84
— of an expression 85
— of competence 476
Normalizer of a formula 91
Noumenal reality 5
Null-class, as a way of speaking 70
Number theory, paradoxes of 6

Object-language 77, 80
Objectivity and science 10
Od and Id, problem of 380
Ontological argument 429
— commitments of discourse 12
Open language 72
Operators, predicate forming 337
Ordinary language 174
Orientational pluralism 12

Paradox of Epimenides 25
— — —, *see also* Liar Paradox; *and* Epimenides *in Author Index*
Paradox-generating propositions 262
Paradoxes 21, 61, 132n
— as traps of logic 74
— of naive set theory 282
— of semantics 227
— of set theory 223, 227
—, truth of 38n
Paradoxical thinking 112, 113
Paranoumena 38n, 39
Particular, concept of 251
Performative arguments 482
— —, inevitably falsified 490
— inconsistencies 483, 486
— —, *see also* Self-reference, pragmati al
Performative inconsistency, irreducibility to contradictions or nonsense 486
— —, problem of determining 492
— —, rejection of is certain 496f
— self-reference 7
— —, *see also* Self-reference, pragmatical
Performatively inconsistent statements 489
— — —, falsified by own utterance 487
— — —, *see also* Self-reference, pragmatic
Phaenomenologia generalis (Kant) 259n
Phenomenological attitude 4
Phenomenology 427
Philosophical argument 443ff
— —, *ad hominem* character of 455
— —, motivation in 449
Philosophical arguments, circularity of all valid 11
— —, feedback in 11
— —, solipsistic isolation of conflicting 451
Philosophical conflict, problem of 448
— research, and goal of extreme comprehensiveness 221
— scepticism 14
— therapy 245
— truth 435
— truths 427
Philosophy 231, 437
—, and self-reference 221
—, first task of 434
—, freedom from convention 436
—, reflexive studies and appli ations 7
—, thought of the second degree 3
—, transcendental 239f
— —, *see* Transcendental philosophy
—, universal metalanguages for 231
Phonetic act 466
Pointlessness 209, 490
— resulting in absurdity of utterances 205n
Positive introspection 348
Possibility, Aristotelian-Megarian view 240
—, conceptions of 240
—, definition in terms of preconditions of valid referring 240
—, Stoic-Diodorean view 240
Possible worlds 278
Pragmatic absurdities 204, 205
— absurdity, as linguistic aberration 206
— criterion of consistency 454
Pragmatic implication 191, 202, 209
— —, and strict implication 211
— —, and verbal conventions 210
— —, as tacit appeal to concept of rational man 198
— —, non-symbolic 192, 193
— —, presupposes system of definitions 193
— —, presuppositions of 208
— —, propositional 195
— —, psychological 195
— —, Strawson's account of 206
— —, symbolic 197f
Pragmatic meaning 210

— paradox 173, 175, 181, 187, 192, 202n, 489
— relation 209
— self-referential inconsistencies 261n
— — —, *see also* Self-reference, pragmatic; Self-refutation
Pragmatical self-reference 7, 8
— —, constructive use of 11
— —, critical use of 9
— self-refutation 173
— —, *see also under* Self-reference
Pragmatics 483
—, formal 167
Pragmatism 10
Pre-reflective experience 14, 15
Precondition of reference 250, 253
Predicate forming operators 337
Predicates 88, 90
—, definition of 338f
—, range of applicability of 132
Presupposition 153, 206
—, and self-reference 224
—, and theory of value 225
—, and truth 154
—, semantical relation of 153
Presuppositional languages 156
Principle of bivalence 154, 155
— — —, metalogical version of 255
— — —, *see also* Law of bivalence
Principle of excluded middle 234
— — —, *see also* Law of excluded middle
Projection 253
Projective forms of reference 254
— meaninglessness 254
Proof theory 35
Propositional attitudes 197, 198, 208, 385, 397n, 403
Propositions 175
— about all propositions 228
—, extensional 278
—, necessarily true 44
— necessitating self-implication 104
—, paradox-generating 262
— possessing denials 229
—, truth of 278
—, universal 278
—, verified or falsified by own utterance 175
Protocol analysis 14n
Protosyntax, self-applied 83
Provability 35
Pseudomenon 26, 27, 35
Psychotherapeutic interventions 6
Putative meaningfulness, problem of 258

Quantification, propositions expressing 277
Quantum mechanics, hidden variable interpretation 14
Quotation marks, metalinguistic use 84
Quotation-name 148

Rational action 195
Real agents 350ff
— numbers, classical theory of 226
Realism 15
— in logic 445ff
Reasoning, default 332
—, intensional 270
Recursive functions 35
— —, general 50
— —, primitive 50, 52
Reductio ad absurdum 22, 23, 41, 42, 132n, 469, 472
Reference, as a ternary relation 251
— failure 138
— of an expression 134
—, precondition of 253
Referenceability 390
Referential consistency 250, 255
— —, as criterion of meaning 247ff
— preconditions 13, 241
Referring capacities 5
— expressions 54
— sentences 251
Reflexibility of languages 72
Reflexive argumentation 13
— phenomena 63
— relations 63
— —, senseless 67
Reflexivity, illegitimate 464
—, and self-reference 6
—, *see* Self-reference
Refutation of monism 494f
— of verifiability theory of meaning 495
Relativistic frames of reference 12
— theory of constitution 17
Representable structures 313
Representation theorem 319-321, 323, 340
— —, proof of 323ff
Restricted propositions 46
Retortion (Isaye) 16
Richard's paradox 6, 43
Rule of signification 206n, 210
— of substitutivity (Leibniz) 396
Rules, semantical 75
—, syntactical 75
Rusell's antinomy 29
— —, *see* Russell's paradox
Russell's paradox 6, 63, 69, 235, 368, 371, 375, 392

S5, as a theory of knowledge 398
Satisfaction 295, 307
Scepticism 10, 482
Self-amendment 457ff
—, contradictions arising from 460
—, *see also* Constitution (legal), self-amendment
Self-applicability 216
—, *see also* Self-reference
Self-application, test of 247
Self-applied protosyntax 83
Self-confirmation 105, 106
Self-consciousness, never reflexive 73
Self-correcting systems 6
Self-describing language 367
Self-epithets 55
Self-implication 104
Self-knowledge 359
Self-nominators 54
Self-reference 153, 213ff, 363ff, 385ff, 464ff, 466
—, and Gödeli ation 217
—, and reflexivity 6
—, and substitution of terms 216
—, and theory of truth 217
—, and translation 213
—, and truth 382
— "bad" 119
—, direct 469
— —, *see* Direct self-reference
—, empirical 122

—, "good" 119
—, genuine 468
—, illegitimate 464
—, in constitutional law 457ff
—, in philosophical argument 7
—, in philosophy 221
—, indirect 122n, 379, 469, 470
— —, *see* Indirect self-reference
—, metalogical 7, 253
— —, *see also* Metalogical self-reference
—, metalogical applications of 13
—, of meaning theory 439
—, paradoxes of 6, 153, 160
—, performative 7, 252
— —, *see also* Self-reference, pragmatical
—, and philosophical methodology 224
—, pragmatical 7, 8
—, pragmatical applications of 9
—, sentential or propositional 252
—, spurious 468
—, uses of 8
—, vice of 467
—, *see also* Reflexivity
Self-referential abilities 8, 363
— argumentation 481
— —, as metaphysical 496
— —, to validate statements about the whole of reality
Self-referential arguments, examples of 493ff
— — — —, Urban's refutation of naturalism 493f
— — — —, Hintikka's refutation of "I do not exist" 494
— — — —, Moore's refutation of "P but I do not believe P" 494
— — — —, Refutation of monism 494f
— — — —, Refutation of verifiability theory of meaning 495
— — — —, *see also under individual names in Author Index*
— —, inevitably question-begging 492
Self-referential inconsistency 242, 481ff
— —, and Cartesian doubt 224
—, and irony 226
— —, as method of philosophical refutation 223
——, at heart of many problems in logic and mathematics 223
— —, paradox-generating 252
— —, pragmatic 261n
Self-referential language 368, 386
Self-referential sentence types 134
Self-referential sentences 119n
— —, determining 133
— —, illegitimate 120
— —, not meaningful 473
Self-referential statements 115
— —, as meaningful 116
— —, as meaningless 485
Self-referential tokens 122, 133
Self-referring assertions 108
— expressions 63
— propositions 471
— sentences 65
— —, *see also* Self-referential sentences
Self-refutation 105, 253, 481
—, pragmatic 173
Self-regulating systems 6
Self-undermining claims 13, 16, 247
— concepts 13
Self-validating assertions 255
— claims 16
— logics 243
— —, to demonstrate transcendental claims 239
— postulates and axioms 243
— statements 115n
Semantic entailment 155n
— paradox 187
— paradoxes 485
— theory 7
Semantical categories 126
— concept of truth 232
— consistency 88
— —, *see also* Consistency
Semantical correctness 137
— —, determining 131
— —, test of 123
— incorrectness 122, 125, 127
— —, *see also* Semantically incorrect sentences
Semantical meta-language 75-77
Semantical paradoxes 115, 120, 147, 223
Semantical sum 79
Semantical systems 75, 83
Semantical truth-definition 80
Semantically incorrect sentences 130
— — —, neither true nor false 128
Semantically null 79
Semantically self-referent language 119n
Semantically universal 79
Semantics, non-translational 75
— of theory of types 232
—, as truth value semantics (Quine) 294
Sense and reference 133
—, as meaning in oblique contexts 266
—, which is a concept 266
Sentence tokens 141, 204
— —, *see* Self-referential tokens
Sentences, classically satisfiable 158
—, falsified by own structure 187
—, necessitation-saturated 158
—, predicating falsity of themselves 187
—, referring 251
—, self-referring 188
—, true only relative to interpretation 154
—, without objects 64
Set theory 32
Significant range 251
Singular terms 291
Skolem's paradox 81
Solipsism 225
Solutions to problems, "discovered" 14
— — —, "invented" 14
Soundness 297
Space 14
—, absolute 15
Spatio-temporal neutrality 183
Stable sets 358
Stack 288
State vectors 278
Statement 202
—, linguistic context of 204
Statements, and belief 208
—, as utterances 202
—, as event-expressions 175
—, as logical expressions 175
—, general 270
— invariant across theories 10
Structural descriptions 78, 79

Subject index, this 503-509
Subjective constitution 4
Subjectivism 11
Substitution, in diagonalization 84n
Substitutivity 390
—, of equals 275
Supervaluation, induced by a set 159
Syntactical meta-language 75
Syntactically self-referent language 119n
Syntaxal 390
System, that can define its own truth 234
Systematic ambiguity 71
— thought, absolute presuppositions of 16
Systems, semantically normal 89

Tarski paradigm 215
— sentence 83, 85n, 92
Tarski's convention T 109
— criterion, unnecessarily restrictive 234, 235
— theorem 83, 84, 93, 94, 371, 412
Tautology 202, 408
TELL, definition of 312
Temporal inference 364
Tenseless language 184
Terms, singular 291
Theorems of formal limitation 6
Theories, about theories 221
—, having no ordinal level (vertical) 222
—, horizontal 222
—, in empirical sciences 221
—, non-ordinal 222
—, of higher ordinal levels 221
—, of ordinal level zero 221
—, self-referentially inconsistent 223
—, vertical 222
Theory, concept of 221
— of argument 7
— of classes, paradoxes 6
— of comprehension 78
— of knowledge 7
— of maximum theoretical generality 226
— of quotation 363
— of types 34, 37, 63, 64, 463, 464
— — —, not necessary 487
— — —, propositions about 38
— — —, ramified 104, 227, 228
— — —, simplified 227
— — —, for types without end 46
Three-valued logic 162n, 262n
— —, matrices 256
Time, absolute 15
— expressions 185
Token oddity 123, 140
Token-reflexive expressions 188
— words 174
Tokens 122
—, *see* Self-referential tokens

Topologically recurved surface 5
Transcendental argumentation 13
— deduction 16, 243
— phenomenology 16
— philosophy 239, 240
— —, metacritique of 239
— —, of science 244
Truth 160, 295, 307
—, absolute 15, 46
—, coherentist theory of 124
—, contingent 422
—, correspondence theory of 124
—, essential 423
—, pragmatic theory of 124
—, semantical concept of 232
—, systems capable of defining their own 49
— value semantics (Quine) 294
Truth-definition 80
Truths, material, necessary 428ff
—, philosophical 427
— which cannot be denied without absurdity 436

Umbrella-words 59
Undecidable systems 95n
Underlying language 345
Unit name 79
Universal language 368
— metalanguages, for philosophy 231
— propositions 278
Universals 43, 278
Unlimitedly false sentences 130
Unrestricted propositions 44, 45
Unsayable things 109
Use and mention 120, 151, 163
Use of words, direct and indirect 265
Utilitarianism, views that oppose 11
Utterance 175, 179, 188, 197, 208
—, inescapable facts relating to 489

Valid arguments, outside formal logic 444
Validity 295, 307
Valuation, classical 157
Veridical paradox 22, 28, 35
Verifiability theory of meaning 440
— — — —, refutation of 495
Vicious circle 463
— — fallacy 37, 70
Virtual implication 261n
— semantics 283

Weil's paradox 68
— —, *see* Grelling's paradox
World structure 296

Zermelo-König paradox 6
Zermelo-Skolem set theory 80