

LOGIC, METHODOLOGY AND PHILOSOPHY OF SCIENCE VI

PROCEEDINGS OF THE SIXTH INTERNATIONAL
CONGRESS OF LOGIC, METHODOLOGY
AND PHILOSOPHY OF SCIENCE,
HANNOVER, 1979

Edited by

L. JONATHAN COHEN

Queen's College, Oxford, England

JERZY ŁOŚ

Warsaw, Poland

HELMUT PFEIFFER

University of Hannover, F.R.G.

KLAUS-PETER PODĘWSKI

University of Hannover, F.R.G.



1982

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM • NEW YORK • OXFORD

PWN — POLISH SCIENTIFIC PUBLISHERS
WARSZAWA

© NORTH-HOLLAND PUBLISHING COMPANY—1982

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission from the publishers

ISBN: 0 444 85423 1

PUBLISHERS:

PWN—Polish Scientific Publishers—Warszawa
and
North-Holland Publishing Company
Amsterdam · New York · Oxford

Sole distributors for the U.S.A. and Canada:

Elsevier North-Holland, Inc.
52 Vanderbilt Avenue
New York, N.Y. 10017

Library of Congress Cataloging in Publication Data

International Congress of Logic, Methodology, and Philosophy of Science, Hannover,
1979. Logic, methodology, and philosophy of science VI.
(Studies in logic and the foundations of mathematics; v. 104)
Includes index.
1. Logic, Symbolic and mathematical—Congresses. 2. Methodology—Congresses.
3. Science—Philosophy—Congresses. I. Cohen, Laurence Jonathan. II. Title. III. Series.
QA 9. A 1157 1979 511.3 80-12713
ISBN 0-444-85423-1 (Elsevier North-Holland)

PRINTED IN POLAND

PREFACE

This volume presents the Proceedings of the Sixth International Congress of Logic, Methodology and Philosophy of Science. The Congress was held at the University of Hannover, Federal Republic of Germany, from August 22 to August 29, 1979, by the Division of Logic, Methodology and Philosophy of Science of the International Union of History and Philosophy of Science. The Congress took place under the patronage of the Niedersächsische Minister für Wissenschaft und Kunst, Prof. Dr. Dr. h. c. Eduard Pestel.

The scientific programme of the Congress consisted of 3 plenary lectures, one memorial, 47 invited lectures and 225 contributed papers, presented orally by the participants. The complete programme is given at the end of this volume.

The general theme of the Congress was

THE ROLE OF MATHEMATICS IN THE SCIENCES

The invited lectures, symposia and contributed papers were presented in fourteen sections, which are listed together with their chairmen and the members of the section programme committees as follows:

- Section 1: *Proof theory and foundations of mathematics*; D. Prawitz;
R. Gandy, G. Mints, C. Parsons.
- Section 2: *Model theory and its applications*; H. J. Keisler;
A. Lachlan, M. Morley, G. Sabbagh, A. D. Taimanov.
- Section 3: *Recursion theory and theory of computation*; Y. L. Ershov;
J. E. Fenstad, Y. Moschovakis, A. Salwicki.
- Section 4: *Axiomatic set theory*; W. Marek;
A. Hajnal, J. Paris, J. Silver.
- Section 5: *Philosophical logic*; B. C. van Fraassen;
V. A. Smirnov, S. Surma, R. H. Thompson.

- Section 6: *General methodology of science*; R. Harré;
V. Lektorsky, U. J. Jensen, G. C. Nerlich.
- Section 7: *Foundations of probability and induction*; I. Levi;
W. Harper, D. Miller, K. Szaniawski.
- Section 8: *Foundations and philosophy of the physical sciences*; G. Toraldo
di Francia; M. Hesse, P. Mittelstaedt, M. Wartofski, R. Wój-
cicki.
- Section 9: *Foundations and philosophy of biology*; M. Ruse;
E. Mendelson, J. Hodge, J. Roger.
- Section 10: *Foundations and philosophy of psychology*; J. A. Fodor;
D. Henrich, O. Chateaubriand, V. Zinchenko, D. Armstrong.
- Section 11: *Foundations and philosophy of the social sciences*; A. Sen;
K. Binmore, S. Kanger, S. Nowak.
- Section 12: *Foundations and philosophy of linguistics*; H. Hiz;
M. Bath, A. Blikle, J. C. Milner, H. Schnelle.
- Section 13: *History of logic, methodology and philosophy of science*;
M. Jammer;
K. Popper, A. Hermann, M. Strauss, R. S. Cohen, L. A. Mar-
kova.
- Section 14: *Fundamental principles of the ethics of science*; D. Føllesdal;
A. Fagot, E. M. Mirsky.

In section 1 a symposium on “The rôle of Constructivity in Mathematics” was organised by R. Gandy. Section 4 was devoted to the memory of the late Andrzej Mostowski. In section 11 the invited lectures were arranged in two symposia, one on “Equilibrium Economics”, the other on “Formal Systems of Rights”. The same was done in section 14, which consisted of a symposium on “Distributive Justice and the allotment of Resources of a Society to Scientific Research” and another one on “Ethical problems involved in Gene Research and Manipulation”. In section 13 there was a symposium on Frege.

These Proceedings comprise the texts of most of the addresses presented by the invited speakers and the main participants in the Symposia. The papers range from reports of new specific results and ideas to more general surveys of recent work. The editors have reverted to the practice of the first four congresses of the DLMPS in publishing the results of the Congress in one volume only. They have tried to express by this procedure the wide scope of the Congress. So the reader will be able to get informed beyond his own field of research as if he had been at the Congress himself.

The Congress was supported by the Deutsche Forschungsgemeinschaft and the government of Lower Saxony. Our thanks are due to these institutions as well as to the supporting institutions of several countries which by their combined efforts enabled so many scientists all over the world to meet at this Congress.

Appended to this preface is a list of the officers of the Division, the Steering Committee coordinating the programme and the Organising Committee of the Congress.

OFFICERS OF THE DIVISION

R. E. BUTTS (Canada)
L. J. COHEN (U.K.)
J.-E. FENSTAD (Norway)
J. HINTIKKA (Finland)
A. A. MARKOV (U.S.S.R.)
P. SUPPES (U.S.A.) President

STEERING COMMITTEE

P. ACHINSTEIN (U.S.A.)
L. J. COHEN (U.K.)
J. Łoś (Poland) Chairman
G. H. MÜLLER (F. R. Germany)
V. N. SADOVSKY (U.S.S.R.)

OFFICERS OF THE DIVISION

R. E. BUTTS (Canada)
L. J. COHEN (U.K.)
J.-E. FENSTAD (Norway)
J. HINTIKKA (Finland)
A. A. MARKOV (U.S.S.R.)
P. SUPPES (U.S.A.) President

STEERING COMMITTEE

P. ACHINSTEIN (U.S.A.)
L. J. COHEN (U.K.)
J. Łoś (Poland) Chairman
G. H. MÜLLER (F. R. Germany)
V. N. SADOVSKY (U.S.S.R.)

ORGANISING COMMITTEE

A. HARMS, Hannover
A. HEINEKAMP, Hannover
G. HESSE, Hannover
M. HOLZ, Hannover
Th. KALUZA, Hannover
I. KÖNIG, Hannover
F. VON KUTSCHERA, Regensburg
G. H. MÜLLER, Heidelberg
A. OBERSCHELP, Kiel
P. PÄPPINGHAUS, Hannover
H. PFEIFFER, Hannover (Chairman)
K.-P. PODIEWSKI, Hannover
J. REINEKE, Hannover
D. SCHMIDT, Heidelberg
K. STEFFENS, Hannover
W. TOTOK, Hannover
G. WILKE, Hannover

LIST OF DONORS

Consumenta Computer, München
Deutsche Bank, Hannover
Deutsche Forschungsgemeinschaft (DFG), Bonn
Deutscher Akademischer Austausch-Dienst (DAAD), Bonn
Deutsche Stiftung für Entwicklungshilfe (DSE), West-Berlin
Gaststätte Beckmann, Hannover
Gaststätte "Bei Bapsi", Hannover
Gaststätte Kaiser, Hannover
Gaststätte Sprengel, Hannover
Gaststätte "Zum Landesknecht", Hannover
Hannoversche Hochschulgemeinschaft
Kali-Chemie AG, Hannover
Land Niedersachsen
Lufthansa
Orma GmbH, Hannover
Rüterbau GmbH, Hannover
Siemens AG, München
Triumph-Adler, Hannover

THE RÔLE OF MATHEMATICS IN PRESENT-DAY SCIENCE

R. THOM

Paris, France

The assessment of the importance of mathematics in the science of to-day is obviously a tremendous task; moreover, as I have been personally involved these last years in some controversies regarding the impact of the so-called catastrophe-theoretic methods, my opinions in this matter cannot claim to represent the general views of the interested scientific community, namely the community of applied mathematicians. I want to make it clear from the outset that, in expressing these views, I may be prejudiced, and I shall be happy if some of these opinions may trigger a useful discussion.

To begin with, let us go back to some historical considerations—philosophers are always fond of the following questions; we may ask: to what extent does mathematics owe its progress to its applications in natural sciences, from mechanics and physics, to biology and social sciences? Roughly speaking, one may divide the advances of mathematics into three types:

- (a) Straightforward applications of known methods or theorems to more exotic or sophisticated material.
- (b) Solving problems arising from the needs of applied sciences.
- (c) Main methodological innovations, associated with introducing new concepts, new axiomatics, and new fields of problems.

Type (c) discoveries are generally due to mathematical reflexion acting on a previously, more or less unconsciously used structure, of empirical or mathematical origin. Examples: The notion of function, the notion of set, the notion of probability.

Advances of type (a) are not going to interest us here. Undoubtedly they form a huge majority of published literature. Their main interest is sociological, namely to select, among the student population, those individuals who are mathematically gifted. The main question to debate here is the

relative proportion of advances of type (b) w. r. to advances of type (c). More specifically: to what extent have great discoveries in mathematics been promoted by problems or methods needed in applied sciences?

A complete study of this question could well motivate a full lecture (if not a complete book!). Here is my—tentative—answer: there is a relative lack of correlation between great advances in mathematics and in mechanics or physics. In most cases of great progress in physics, the body of mathematical tools anticipated the physicist's needs, and it very seldom happened that the mathematician had to create a new theory to satisfy the needs of the physicist. Here are some examples:

Kepler's use of the theory of conics due to Apollonius;

Galileo's use of elementary algebra (linear and quadratic functions);

Einstein's use of Lie group theory for special relativity;

Einstein's use of tensor calculus for general relativity;

Schrödinger's and Heisenberg's use of Hilbert space for quantum mechanics.

As counterexamples one may quote:

Fourier's analysis motivated by wave-optics;

Distribution theory motivated by Dirac's δ -function. As an ambiguous case:

- Newton's differential calculus motivated by his mechanics.

The last case is ambiguous, because Newton had many forerunners in his Calculus (Archimedes, Cavalieri and Pascal being among the most famous...), moreover, he did not have a general concept of a function. Here the question boils down to the following: is the notion of instantaneous velocity, the obvious source of the notion of derivative, a truly scientific or an intuitive concept? I am inclined to say that is an *intuitive* concept.

Finally, as the only clear-cut case of a mathematical theory inspired by physics, I find Fourier's series and transformation and its extension into functional analysis. The theory of P.D.E. (partial differential equations) was of course largely motivated by mechanics and physics, with the result that it was—and still is—largely confined to linear theory: nonlinear theory had to wait for Cauchy and Kowalewski for a start...

My main thesis is also justified by history: a good deal of mathematics (Greco-Latin plus Arabian mathematics) existed before modern science was born in the XVIIth century. And much can be said in favour of the idea that modern science originated with the slow maturation, in people's minds, of the mathematical concept of a function, which was only made

explicit in 1695 by Leibniz. In more recent times, i.e., since quantum mechanics, there has been a growing gap between theoretical physics and the main stream of mathematics. Quantum field theory needed, for its conceptual justification, a great deal of mathematical elaboration in functional analysis. Mathematicians were rather slow to enter this strange type of mathematics and, despite recent advances, the situation still cannot be termed satisfactory.

Another permanent source of inspiration for mathematics is, of course, the notion of chance—conceptualized as probability. The theory of games is a very vast domain, which still awaits mathematical elaboration. The same can be said of control theory. With statistics and data analysis, we are entering a field which obviously satisfies very urgent technological needs, but in which the underlying theory is relatively undeveloped. It seems that in such fields the practice of applied mathematics mainly consists of using a certain number of tricks, or recipes, the theoretical justification for which is frequently lacking.

The general conclusion of the above discussion reflects nothing but a well-known triviality in the philosophy of science: namely, that it is difficult for us to discover facts in our experience if we do not have for these facts an *a priori* image, a mental model which enables us to simulate internally the external phenomena. And the real importance of mathematics is that it furnishes us with models which could not be constructed by natural language alone.

The rôle of mathematics in science, is, of course, closely related to its philosophical, ontological status. Hence it may not be entirely out of place to give some side remarks about the philosophical nature of mathematics. (see Note 1).

I consider “logicism” to be a completely false view of mathematics. The foundations problems—which are going to occupy so many of you during this congress—are, for the working mathematician, by no means fundamental. I believe that extensional, set-theoretic logic may give rise to nice technicalities, but has led the philosophy of mathematics into what our British neighbours would call a “cul-de-sac”. In everyday language, the extension of a concept (think for instance of the concept “red”) is always a “fuzzy” set, with no fixed boundaries; hence any true logic has to be intensional and cannot be extensional unless it deals with concepts of an artificial nature, whose extension can be generated by a constructive procedure. While this may be the case for many mathematical concepts, the logic used by a mathematician while he is working is fundamentally in-

tensional and always implies the “meaning” of the concept. I have described the situation with the somewhat provocative motto: “Tout ce qui est rigoureux est insignifiant” (cf. References). Of course, mathematics, despite its rigour, is not meaningless. The meaningful character of mathematics arises from its dealing with the geometric continuum, with the intrinsic spatial character of mathematical objects. More specifically, all fundamental theorems in mathematics deal with local-global problems. Let me quote here a list of theorems or methods which play a fundamental rôle both in pure mathematics and in the applications of mathematics, especially in physics.

(a) *Taylor's formula.* Let $f(x)$ be a real-valued function defined on $a \leq x \leq b$, provided with k continuous derivatives $f^i(x)$, $0 \leq i \leq k$; then, near any point x_0 of the segment $[a, b]$ we have the formula

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \dots + f^{(i)}(x_0)(x - x_0)^i/i! + \dots \\ \dots + a(x)(x - x_0)^s/s!, \quad 1 \leq s \leq k,$$

where $a(x)$ is a function of x which has $k-s$ derivatives and is bounded in modulus by $f^s(x)$ on $[a, b]$.

(b) *The implicit function theorem.* Let $f(x, y)$ be a function of two variables (x, y) defined on a domain D of the plane Oxy with its partial derivatives continuous up to the order k . Let (x_0, y_0) be a point of D where $f(x_0, y_0) = 0$, $f_y(x_0, y_0) \neq 0$; then there exists a function $y = g(x)$, provided with k continuous derivatives, defined near x_0 , such that $f(x, g(x)) = 0$ near $x = x_0$.

(c) *Stoke's Theorem.* If a is a closed differential form defined on a simply-connected domain V of the Euclidean space, $da = 0$, then there is a global real-valued function $V: U \rightarrow R$ such that $a = dV$.

(d) *Analytic continuation.* Any germ of analytic function f around x_0 may be extended throughout the whole holomorphy domain of f , where f is uniquely defined.

Theorems (a), (b) define restrictions from global to local; but the local form is somewhat more specific than the global form one started with. By iterating such local restrictions, one may get useful global information (for instance, about the topological structure of the set $f(x, y) = 0$ in case (b)).

Theorems (c) and (d) concern the inverse arrow local \rightarrow global; observe that theorem (a) defines the localization procedure inverse of process (d);

process (c) has itself an inverse, as

$$a = dV \quad \text{implies} \quad da = 0 \quad (dd = 0).$$

Now it happens that these theorems have a bearing upon the way mathematics can be applied in science. There are two fundamental objectives for science. The first is pragmatic: to increase the power of mankind over its environment in such a way as to ensure the survival (and if possible the expansion) of the human species. The second is speculative (theoretical): to get a better understanding of the world around us. At first sight, these two objectives may seem nearly identical. In fact, they are not: for it is quite possible to discover a very useful recipe the efficiency of which may completely escape our understanding (the history of pharmacology is full of such examples); and it may well be that pure understanding of a hopeless situation may be so complete as to discourage any attempt to remedy it. Acting and understanding are to be considered as the two poles of a continuous spectrum of scientific activities: of course, understanding does not necessarily harm practical efficiency, and control by practice (experiment) is a good check of the truth of the understanding...

But the fact remains that these two poles of scientific enterprise require fairly different types of tools. And this is the case in mathematics. The pragmatic pole is bound to the possibility of prediction. And any theoretical scheme, capable of directing human activities, has to convey a procedure of spatio-temporal localization. For if we want to act, we always have to act in a specific place (in a localized system) at a given time. Moreover, human activities always try to realize what does not occur naturally; hence, we always want to go beyond our natural capabilities: mankind's recent history testifies to the tremendous increase in space, in time, in energy, in velocities, which are now within our reach. As a result, all these attempts at extending human power rely on mathematical tools permitting the extension of data from a smaller domain to a larger one, hence on procedures of the type local → global. Theorem (c), Stoke's theorem, is in itself not sufficient for that; for the function V is defined only up to a constant, and this arbitrariness may destroy the possibility of quantitative prediction. (What the physicists call the now so fashionable gauge theory is nothing but a set of procedures used to remove such indeterminacy.) Hence the basic tool for extending data remains analytic continuation (d).

This is in fact what happens: the best possibilities of prediction are always associated with analytic equations whose solutions may be extended by continuation. (Think, for instance, of the computation of a satellite's

orbit, defined by Newton's law.) But this possibility of accurate prediction requires first that the underlying substrate space (where the extension is to take place) has a "natural" analytic structure. Second, even if this condition is satisfied, analytic continuation by itself may be a rather poor way of extrapolating data. For approximating empirical data up to ε by an analytic function f on a compact set K does not suffice in general to determine the holomorphy domain of f . If we want to have a good control of the extrapolation procedure, we need an underlying theory which restricts the possible form of the function f . This is the case in fundamental physics: the substrate space is then the Minkowski space, a homogeneous space of a Lie group, and hence provided with a natural analytic structure. And physical theories (relativity, quantum mechanics) allow us to define the localization of interesting elements (particles) by analytic functions of a specific form given by the theory. Here we are at the root of the so-called miracle of physical laws, which enabled science to produce its most powerful procedures for prediction. To what extent can this miraculous situation be found in other cases than fundamental physics? We shall see later that, in all likelihood, the miracle of physical laws is unique and isolated; hence our capabilities for quantitative prediction are severely limited to domains close to physics and mechanics.

Of course, the requirement of having a substrate endowed with a natural analytic structure may not seem very strong. In statistics, for instance, the probability of the occurrence of a given (qualitatively defined) event defines a real axis R , which has a natural analytic structure. And we might be tempted to use this structure for extrapolation, especially because all the standard laws of probability (Poisson–Gauss laws) are defined by analytic functions with respect to this structure. I have expressed elsewhere (THOM, 1979) all the doubts aroused by such procedures; for if we do not have *a priori* notions of the mechanisms generating the events studied, there is no reason to believe that the observed probability distribution be analytic as a function of control parameters of spatial character.

Let us now pass to the other pole, namely the need for understanding. Against the empiricist attitude—so frequently found among modern scientists—it may be necessary to repeat that understanding things may be more than a superfluous luxury. It does not suffice to know how things behave; for that, it suffices to make experiments and to store the results in a computer's memory. We also need to understand how things behave; this is the only way for us to decide which experiments have to be made and which do not. Here we find the other rôle of mathematics in science, namely *model*.

building. By comparing the results of a model—either quantitatively or qualitatively—with experiment, we are able to test the validity of the model. Of course, the methodology involved in model building and testing is an intricate one.

Generally, model builders are quite happy to find a relatively good quantitative agreement of their model with experiment. This might mean very little, as long as one does not have a uniqueness proof for the model—at least among a given class of models. Catastrophe theory methodology—in my sense—does not insist on quantitative agreement, but tries to exhibit the simplest models compatible with a given experimental morphology. This is an entirely different viewpoint, directed less to the study of the phenomenon itself than to its surroundings, its genesis and disappearance...

Which are the mathematical tools needed for this interpretative task of mathematics? Facing any given process, the first thing to do is to isolate the stable morphological accidents appearing in the process. We have first to look for *singularities*, to classify—identify them—and, at a later stage, to try to reconstruct the global process from its singularities as generating elements. For instance, in gas kinetics, one considers first one-dimensional singularities (the moving molecules), and later zero-dimensional (punctual) singularities, and the collisions between two (or more) of them. This second stage of reconstruction, if we want to make it quantitative, requires some form of analyticity of the global evolution with respect to its generating singularities. (For instance, a holomorphic function is defined up to a unit by its zeros.) As a result, quantitative checking of a model makes sense only if the underlying analyticity hypothesis can be justified *a priori*.

The analyticity postulate “An analytic function is defined by its singularities” is analogous to the logistic postulate “The meaning of a concatenation of elements in a formal system depends only on the meanings of its elements”. In real linguistics unlimited iteration of operations (competence) is limited by performance. In the same way, analyticity may be limited by loss of differentiability (for simple differentiability of finite order may suffice to define singularities). It would be of the utmost importance for applications of mathematics to determine the “nuclei of analyticity” in which reality may be decomposed. May be the boundaries between them, where one has only differentiability without analyticity, would correspond to qualitative differences between different disciplinary fields.

We have finally arrived at the general conclusion: there are basically two types of applications of mathematics in science:

- (1) A precise, accurate quantitative way, in which one is able to make precise quantitative predictions. This is the situation in mechanics and fundamental physics.
- (2) An interpretative—hermeneutic—way, in which one tests hypotheses and models.

This second way generally does not bring certitude; probabilistic methods are those for which—in ways which are frequently themselves subject to doubt—one is able to measure quantitatively our ignorance or our doubts.

It is well known that the efficiency of quantitative formalism rapidly decreases as one goes from fundamental physics to macroscopic-phenomenological physics, to chemistry, and later to biology. In the social sciences, quantitative formalism is used only for statistics. But an interpretative use of the second type remains possible—and frequently useful. But it cannot bring certainty... This is perhaps the reason for the ambiguous status of applied mathematics in the sociology of science. Disciplines like biology, social sciences (sociology...), which do not have precise quantitative laws, are fond of mathematical techniques, because it brings them the prestige of accurate predictive disciplines like astronomy or physics... And computer industry has any reason to make us believe that no field of reality may escape the validity of quantitative formalism.

It is often believed that introducing mathematical techniques in the methodological apparatus of a discipline means fundamental progress for that discipline. Such an opinion is illusory; the use of mathematics in any discipline whatsoever always has to be assessed in its proper place according to pertinent criteria, derived both from the actual situation and from general epistemological considerations.

I think it extremely important for the ethics of applied mathematics to have the general opinion (among scientists and laymen) realize that there is a *non-exact use of mathematics*, which may nevertheless have some usefulness, whether practical or theoretical. The unconscious wish to hide this fact has led the corporation of applied mathematicians into a kind of intellectual marasme. First of all, professional applied mathematicians—at least in the West—appear to be the product of a counter-selective procedure. Those mathematicians who apparently are not gifted enough to become pure mathematicians are directed towards applied careers. It seems obvious, however, that it is much more difficult to be an original applied mathematician than a honourable pure mathematician. For the applied mathematician has to know and thoroughly understand pure mathematics,

and also has to know—and understand that part of reality which he is studying. Hence he must be a mind of rare quality, able to join concreteness and abstraction. This is perhaps the reason why I think it is fair to say that no applied mathematician of great stature has appeared since the death of John von Neumann (see Note 2). It would also be worthwhile to convince pure mathematicians that the study of experimental disciplines could bring a great deal of new material to pure mathematics itself...

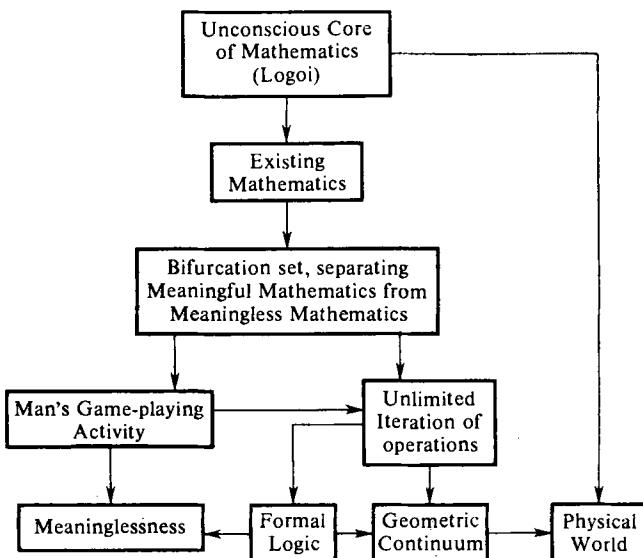
In 1880, Henri Poincaré introduced a completely new point of view, in the theory of differential equations, by substituting for the research of explicit quantitative solutions the qualitative—topological—study of the global “phase portrait” of the system. Catastrophe theory—in my sense—advocates the same viewpoint for modelling reality. Models so constructed interpret morphological qualitative differences as jumps between “attractors” of internal dynamics. Such a method tends to bridge the gap—open since Galilean dynamics—between quantitative formalism and everyday language description. Because of this mixed structure, catastrophe-theoretic models may appear inferior to both types they join: inferior to standard quantitative models, because they lose numerical accuracy, and inferior to verbal description, because they do not have all the richness and subtleties of conceptual thinking.

It may be worthwhile to remind the reader about the motivations for “geometrizing meaning”; there is little doubt that problems considered in philosophy are of general interest to all mankind—much more important, in fact, than many scientific problems. Hence their statements and possibly their solutions should be amenable to an intersubjective understanding at least comparable to the common consensus achieved for scientific questions. There are many reasons to believe that no sharp distinction between science and philosophy can be drawn. First, many apparently purely scientific questions—such as the epigenesis of space in the child’s development—cannot be attacked without some philosophical background. Conversely, in social sciences, the use of natural language, and of purely conceptual thinking, leads to a way of reasoning which is very frequently extremely intricate and subtle, especially when the author introduces his own terminology. If we are able to transform such reasoning into purely geometric (topological) picture, then we may ensure to a large extent its objectivity; by using the “distantiation” effect of geometric representation, we are able to escape the fascination of meaning, to break the hermeneutic circle which has kept imprisoned so many social science thinkers.

Why should scientific techniques be completely powerless with respect to philosophical questions? I strongly hope that these models will open a new way of thinking in science, which may, by its speculative character, trigger a revival of nature philosophy.

Notes

1. For a global description of the ontological status of mathematics see the following diagram:



the “core of unconscious” structures, unknown to us, creates mathematics in the human mind, and also acts as organizational morphological entities in the real physical world. Man’s game-playing activity plays on this material and thus creates logic and mathematics by the unlimited iteration of algebraic operations. Only those sequences which can be naturally embedded in the geometric continuum become meaningful. Others end in meaninglessness.

2. Such an abrupt statement needs perhaps some qualification. In the Western countries, one should take into account the sad fact that John Nash’s mathematical career was prematurely interrupted. More recently, some bright applied mathematicians (Hénon, Lorenz, ...) have initiated the revival of iteration theory, one of the most active fields of the present-day dynamics: a theory which started with the works of Fatou–Julia in 1920

and then fell into oblivion. In the Soviet Union and the European socialist countries, the distinction between pure and applied mathematics has never been strictly enforced as it is in the West. As a result, in both disciplines there have been famous names, such as Kolmogoroff, Pontrjagin, and more recently Maslow.

References

- THOM, R., 1973, *La Science Malgré tout*, Encyklopedia Universalis Organum. vol. 17, pp. 5–10 (Paris)
- THOM, R. 1979, *Mathematics and Scientific Theorizing*, Vol. Scientific Culture in the Contemporary World (UNESCO, Scientia Milano)

“SCIENTIFIC” HISTORY AND TRADITIONAL HISTORY*

ROBERT WILLIAM FOGEL

Harvard University

Although in some respects the verbal clash between “scientific” and traditional historians is as loud as ever, the interpenetration of their opposing modes of research is now quite well advanced. The long period of cultural warfare has turned some of the partisans into determined enemies, but they are few in number. The clash is not rooted in irreconcilable ideological positions or in territorial struggles, even though these may be involved incidentally. It turns on more subtle and more complicated differences over research agendas, methodology, and style. Reflection on the substance of these differences suggests that the affinities and complementarities between “scientific” and traditional historians are more important than the differences. Nearly a quarter of a century has elapsed since the revival of the conflict. Placing the points at issue in an historical perspective not only serves to clarify past events but also suggests the nature of the emerging synthesis.

* I am thankful to Professor Jerzy Łoś and the other members of the organizing committee of the Sixth International Congress of Logic, Methodology, and Philosophy of Science who commissioned the paper. Exchanges with Per Boje, Noel Butlin, Stanley L. Engerman, David H. Fischer, Morton Keller, Alan Kulikoff, and Donald S. Lamm clarified many difficult issues. I benefitted from the discussion of a late draft of this paper at a seminar of the fellows of the Warren Center for Studies in American History as well as from comments and criticisms by William O. Aydelotte, John Clive, A. W. Coats, Thomas C. Cochran, François Crouzet, Philip Curtin, Lance E. Davis, Mary Dobson, Jon Elster, G. R. Elton, Roderick Floud, Claude Fohlen, Robert Forster, Gerald Friedman, David Galenson, G. R. Hawke, Samuel P. Hays, David Herlihy, Patrice Higonnet, Eric Hobsbawm, J. R. T. Hughes, J. Morgan Kousser, Frederic C. Lane, Peter Laslett, Seymour Martin Lipset, Donald McCloskey, Douglass C. North, Barbara Rosenkrantz, W. W. Rostow, Arthur Schlesinger, Jr., Roger Schofield, Joel Silby, Daniel Scott Smith, Kenneth Sokoloff, Lawrence Stone, Cushing Strout, Robert Swierenga, Stephan Thernstrom, Charles Tilly, Harrison White, John Womack, and E. A. Wrigley.

The quest for scientific history

By the early decades of the nineteenth century, if not sooner, it was clear that scientific and technological discoveries were radically transforming the mode of production and altering long-established habits of life. The industrial revolution was, on final analysis, the gift not of liberal kings and democratic politicians, but of practical men who saw how to transform these discoveries into new methods of production. In that incredible burst of inventive activity and entrepreneurship that did so much to transform the nations of Europe and North America, it seemed as if science had no bounds. During the first half of the nineteenth century space and time were conquered by steamboats, railroads, and telegraphs. The prices of one-time luxury commodities declined so rapidly that they were transformed into necessities, not only for the rich and the well-to-do, but also for the common people. And a wide array of commodities and services came into being or were popularized.

It was quite natural that those who pondered the nature of human affairs should have been deeply impressed by the transformation that resulted from the applications of science to production. If the application of science to production could transform industry, could not the application of science to social and political relationships transform these as well? The question was all the more urgent, because Europe and America had already entered an age of social and political revolution and because long-standing ideologies were coming under increasing challenge. Stimulated by the thought of Enlightenment thinkers, Malthus sought the laws that governed the growth of populations; Ricardo sought the principles that governed the distribution of income; Morgan began the study of systems of kinship; and LePlay sought to establish regularities in the organization of families. From their work and the work of like-minded scholars, there arose such disciplines as economics, demography, anthropology, and sociology. From their work, as well as the work of biologists, there also arose a new subfield of mathematics called statistics. And the collectivity of these fields, which sought to use the methods of the natural sciences to study the social behavior of human beings and to derive law-like statements that governed their behavior, became known as the social sciences.

History was also deeply influenced by the natural sciences, but the lines of influence were complex and somewhat contradictory. To some historians being scientific meant being objective, and objectivity required a rejection of conceptions of history that emanated from moral philosophy. Objectivity,

as defined by Leopold von Ranke, required, above all, a deep immersion in the primary sources and the subjection of these sources to intense internal and external criticism. This quest for the authenticity of evidence, which involved techniques developed in classical philology, including paleography and diplomatics, led to the heavily footnoted monograph as the foundation of historical research. It also led to an emphasis on the unique and particularistic nature of history and to a rejection of the applicability of law or law-like statements to the study of history.¹ As Charles H. Hull put it, "the ultimate units with which the historian deals are not atoms, or any sort of instrumental abstractions, whose individual differences may be ignored, but they are men and the deeds of men." Men and their deeds, he held, were "too complex and too variously conditioned to be subject to the concept of general law" (HULL, 1914, p. 35).

This is the point on which advocates of scientific history were most deeply divided. Many historians during the late nineteenth and early twentieth centuries, deeply influenced by the revolutionary discoveries in biology, especially the work of Darwin, and by the integration of biology with physics through the laws of thermodynamics, began to propound the view that "history represented a continuum with the universe of nature, and like nature, was ... governed by law." In his presidential address to the American Historical Association, Henry Adams spoke with awe about the "immortality that would be achieved by the man who should successfully apply Darwin's methods to the facts of human history." Such a breakthrough seemed at hand to Adams who contended that "four out of five serious students of history have in the course of their work, felt that they stood on the brink of a great generalization." That was in 1894. Fifteen years later, the breakthrough was yet to be achieved and no longer seemed imminent. When George Burton Adams delivered his presidential address he warned against efforts to discover "the forces that control society" or "to formulate the laws of their action," calling them the "allurements of speculation." "The field of the historian," he said, "is, and must long remain, the discovery and recording of what actually happened."²

And so the majority of the profession removed themselves from the effort to establish empirically-warranted laws of human behavior, and appeared

¹ Recent discussions of Ranke's methodology and influence include those of IGGERS (1962), GAY (1974), and KREIGER (1977).

² SAVETH (1960, p. 4), H. ADAMS (1894, p. 19), G. B. ADAMS (1895, p. 236). The rise and decline of scientific history during the late nineteenth and early twentieth century are described in HOLT (1940), SAVETH (1960), and HIGHAM *et al.* (1965).

content to leave such speculative enterprises to the social sciences. To some historically-minded social scientists this seemed to be a reasonable division of labor. John William Burgess, for example, told historians that their task was the "true and faithful" recording of "the facts of human experience," and that political scientists would put these facts "into the form of propositions" or "principles" of political behavior. (BURGESS, 1896, pp. 209-210.)

Historians, of course, generally refused to accept such a narrow definition of their function. Many economic historians, for example, still aimed to produce law-like statements, not by the deductive process that animated Marshall and other economic theorists, but through an inductive process, which began with the systematic collection of the facts of economic history. (See FOGEL, 1965, and the sources cited there.) And if the main body of historians, concentrated largely in political history, gave up the search for behavioral laws *per se*, they did not give up looser and what seemed to them less pretentious forms of generalization.

The triumph of traditional history

The idea that there could be a special historical approach to the interconnection of facts was given systematic expression by Johann Gustave Droysen who, drawing heavily from the work of Wilhelm von Humboldt, sought to define a process called *Ideengeschichte* or "imaginative understanding." As Jurgen Herbst summed it up, historical truth "consisted not only of facts but of ideas—invisible elements of history 'which relate the fragments to each other, put single phenomena in their proper perspective, give form to the whole'." To "fuse the visible facts with the invisible ideas" required imagination. In exercising this imagination "the historian resembled the poet, except that it was the historian's special duty to subordinate his imagination to an absolute fidelity to past actualities" (HERBST, 1965, pp. 108-111).

According to Droysen, historical thought was a well-defined "epistemological category" that was "distinct from those of the physical sciences and of speculative philosophy". Toward the end of the nineteenth century and early in the twentieth century philosophers such as Dilthey and Croce, and later Collingwood, sought to explore this new epistemological category and thus gave rise to the subfield called "the analytical philosophy of history." They held that historical agents could not be viewed as "mere pieces of observable 'behavior', reducible to (or explicable in terms of) purely physical items." It followed that the "essential task of the historian is to 'rethink'

or 're-enact' in his mind the deliberations of historical agents, thereby rendering intelligible the events with which he has to deal in a way that finds no parallel in the physical sciences" (HERBST, 1965; GARDINER, 1968, p. 430; cf. H. HUGHES, 1958).

With such philosophical underpinning, historians tended to move further away from the notion that natural science provided a model for historical research. In America the climax of the process of disenchantment was probably reached in the 1930s with the presidential addresses of Carl Becker and Charles Beard, which Cushing Strout has described as a pragmatic revolt against scientific history and historical positivism. Becker and Beard denied that the "historian's account of the past" could be "genuine knowledge" except "to a very limited degree." An historical account, they argued, was "fundamentally a temporary appraisal, based on the historian's interests and values, which are themselves conditioned by his particular time, circumstances, and personality." While it cannot be said that the majority of historians adopted this view of their craft, many were influenced by it, and during the 1930s, 1940s, and early 1950s there was a continued weakening of the earlier identification with the natural sciences. This was true even in the collection and appraisal of evidence, and historians gradually shifted from the natural sciences to law as their model for the rigorous evaluation of evidence (STROUT, 1966, p. 9. Cf. HIGHAM *et al.*, 1965, Chs. 3 and 4).

The dominance of the legal model is quite evident in the standards for the appraisal of evidence set forth in the 1954 edition of the *Harvard Guide to American History*.³ The Harvard historians first warned of the need to determine carefully the actual meaning of words in particular contexts. Some words have a drastically different meaning today than they did in the past, while other words, even at a given moment of time, have a different meaning for one group than for another. "Property," they pointed out, "meant one thing to John Locke, another to the American Liberty League." They also emphasized the need to be sensitive "to irony, satire, epigram, literary flourish, overstatement, understatement, and the whole human range of nuance, inflection, and exaggeration."

Once satisfied that he understands what the witness is saying, the historian must then consider whether the witness was in a position to know what he was talking about; then whether, if the witness was in

³ The quotations in this paragraph, including the indented quotation, and the next one are from HANDLIN *et al.* (1954, pp. 24-25).

that position, he had the skill and competence to observe accurately; then whether, if he knew the facts, he would be inclined to represent them fairly, or whether circumstances—emotional, intellectual, political—might incline him to emphasize some aspects of an episode and minimize others. Many motives, worthy and unworthy, deflect or distort observation: national patriotism, class conditioning, political partisanship, religious faith, moral principle, love, hate, and survival.

To get at the truth, the historian must cross-examine the witness. The Harvard historians provided sensible guides to this process. They pointed to the need to take account of the “general character” of each witness, but warned about “the logical fallacy of confusing the origin of the story with its value: the chronic liar may bear true witness on a particular occasion, and the high-minded man may get hopelessly muddled.” They also warned about the “careless” errors in letters and the “premeditated” entries in diaries, about the faulty or self-serving recollections of aged men, and the easy acceptance of “poetic fancy ... as historical fact.” Because it “is all too easy to poison the bloodstream of history,” they emphasized the need to find independent corroboration for any proffered fact:

Like treason in the Constitution, a historical fact ideally should rest “on the testimony of two witnesses to some overt act, or on confession in open court”. But sometimes, alas, there is but a single witness; or, if there are two, and of equal competence and probity, their versions may be in head-on collision. Charles Evans Hughes told his biographer that he had recommended Robert H. Jackson as chief justice of the Supreme Court; President Truman’s firm recollection is that Hughes recommended Fred M. Vinson; no documentary evidence survives: how to solve the insoluble conflict?

A judge and jury, indeed, would go mad if they had to decide cases on evidence which will often seem more than satisfactory to the historian. But there is no escape; the historian, if he is to interpret at all, will try and convict on evidence which a court would throw out as circumstantial or hearsay. The victims of the historical process have to seek their compensation in the fact that history provides them with a far more flexible appellate procedure. The historian’s sentences are in a continuous condition of review; few of his verdicts are ever final.

I have dwelt on standards of evidence, because it is central to the distinction between what I have called “traditional” history and the brand of

"scientific" history that is now successfully challenging it. Before proceeding to a comparison of these two modes of historical research, it is first necessary to consider the subject matter of traditional history.

During the nineteenth century historical research, not only in the United States, but also in England, France, and Germany was focused on politics. The words that Herbert Baxter Adams inscribed on the wall of his seminar are so famous now, that they are a cliché: "History is Past Politics and Politics are Present History." The aphorism was not original with Adams, who took it from Edward A. Freeman. Nor was preoccupation with politics new. From the beginning of the craft, historians had focused on the state, the church, and other organs of power; and the authors often had been persons who held positions of power or were close to those who held them. During the era of the absolute rule by divine sovereigns, historians were often figures at court who sought to glorify the deeds of their patrons. After the onset of the age of revolution, politics continued to remain the focus of historical writing, but then the historians celebrated the accomplishments of the revolution and explained the reasons for the downfall of the old regime. The rise of scientific history in Germany, and its subsequent spread through Europe and America, further accentuated the emphasis on politics. For state papers were the most carefully preserved and easily accessible of the primary sources that Ranke and others had made quintessential (H. B. ADAMS, 1895, p. 67; BARNES, 1962, Chs. 4-10; MARWICK, 1976, Ch. 2).

Early in the twentieth century strong movements developed in Europe and America for histories that included, but transcended, politics. The origins of these movements can be traced to such earlier scholars as Giambattista Vico, Lord Macaulay, and Jules Michelet, whose broader, more social conception of history represented an alternative to the dominant political tradition. A shift in the balance began to become evident about the time of World War I or shortly thereafter. In the United States James Harvey Robinson called for "a new history" that would embrace "every trace and vestige of everything that man has done or thought since first he appeared on the earth." Such history, he continued, would not only "follow the fate of nations" but also "depict the habits and emotions of the most obscure individual." In France the movement for "total history" revolved around the journal *Annales*. Perhaps more than any group in any other country, the *Annales* group eagerly drew upon the full array of ideas and approaches generated by the social sciences, including the development of a sociological approach to politics. They have also displayed great

methodological flexibility, so much so that in recent years such prominent members of the group as Emmanuel Le Roy Ladurie and François Furet have been in the forefront of French scholars seeking to apply quantitative methods to the study of history. Collectively, the members of the *Annales* group have explored nearly the full range of approaches now practiced in the study of history, although some have a greater affinity to traditional approaches while others have a greater affinity to the new brand of "scientific" history.⁴

American historians were somewhat ambivalent regarding what they should or should not take from the social sciences. At first economic ideas were the most easy to incorporate. Frederick J. Turner laid great emphasis on the effects of the vast supply of land in the shaping of American democracy; Charles A. Beard attempted to show how differing economic interests influenced the constitutional process and the course of political conflict; and U. B. Phillips drew on the economic theory of capital to explain the course of slavery during the antebellum era. Toward the end of the interwar period and beyond, sociological and anthropological ideas seemed increasingly relevant. Richard Hofstadter showed how pervasive the influence of social Darwinism had become in the thought of the late nineteenth century, while Oscar Handlin explored the opportunities and obstacles to social mobility among immigrants to Boston.⁵

What I mean by "traditional" history, then, is the type of history that was described in the 1954 edition of the *Harvard Guide* and that was practiced during the 1930s, 1940s, and 1950s, by its authors and by such other distinguished historians as C. Vann Woodward, Kenneth M. Stampp, Allan Nevins, and Richard Hofstadter in the United States; by R. H. Tawney, G. M. Trevelyan, Herbert Butterfield, J. H. Plumb, and G. R. Elton in Great Britain; and by Marc Bloch, Lucien Febvre, and Fernand Braudel in France. The traditional historians aspired to portray the entire range of human experience, to capture all of the essential features of the civilizations they were studying, and to do so in a way that would clearly have

⁴ ROBINSON (1912, p. 1). For discussions of "the new history" see BARNES (1962), BRINTON (1939); HIGHAM *et al.* (1956). For discussions of the *Annales* group see AYMARD (1972), BAILYN (1977), FORSTER (1978), HEXTOR (1972) and STOJANOVICH (1976); on their use of quantitative methods see FURET (1971), FURET and LADURIE (1970), LADURIE (1979).

⁵ See G. R. TAYLOR (1975), HOFSTADTER (1968a and 1955), BENSON (1960), FOGEL (1970), and HANDLIN (1941). On the turn toward sociology and anthropology see LE GOFF (1971), HOFSTADTER (1968b), SAVETH (1964), and STONE (1977b). Cf. H. HUGHES (1971), GILBERT (1971).

relevance to the present. They were continually searching for "synthesizing principles" that would allow them to relate in a meaningful way the myriad of facts that they were uncovering. This search led increasingly to generalizations that were emanating from the social sciences. While some found that the economic interpretation of history provided the best "conceptual framework on which to order the whole," others preferred the sociological or cultural conceptions. But the most common tendency was to be eclectic. Historians increasingly took from each of the social sciences those ideas that could add power and depth to their analyses, without committing themselves to one all-embracing view of human behavior or historical evolution. An intuitive notion of "imaginative understanding" or "historical imagination" remained the basis for overall thematic integration (HOFSTADTER, 1968b, p. 8).

Despite the large and increasing role of social-scientific thought in their work, many traditional historians have been wary about the generalizations produced by the social sciences and highly selective in what they have accepted. "Too careful an ear cocked for the pronouncements" of social scientists, G. R. Elton warned, is "liable to produce disconcerting results," partly because "in the social sciences fashions come and go with disconcerting speed." It is not only the poor control of the quality of generalizations that has led Elton and other traditionalists to resist the scientific embrace, but a belief that history is an autonomous discipline with standards of scholarship better suited to the tasks of historians than those developed elsewhere. Much turns on the relative importance of the general and the particular. Elton characterized history as "ideographic", that is it particularizes, and not 'nomethetic', that is, designed to establish general laws". He doubts that the particulars which historians study will ever "become numerous enough for statistical generalizations from them to be valid," and he is convinced that historians must treat "facts and events (and people) ... as peculiar to themselves and not as undistinguishable statistical units or elements in an equation." While stressing the autonomy of history, Elton and other traditional historians shun exclusiveness. They recognize that historians can learn from the "social scientist's precision, range of questions, and willingness to generalize," and "may often be well advised to count heads." But they are quick to add that such matters "can never be more than a small part of the whole enterprise." While Elton's views may come close to representing the central tendency of traditional historians, the range of their attitudes toward the social sciences is quite wide, and those who have been more radical in methodology, such as Handlin and

Braudel, have done much to pave the way for the new brand of "scientific" history.⁶

There are grounds for an argument against classifying scholars as different as Elton and Tawney or Nevins and Handlin into a single category called "traditional historians." Not only is the range of approaches among traditional historians wide, but the gradations in approach are very fine. One could propose a variety of useful ways of breaking the large class into a number of smaller ones. For some purposes, for example, one ought to distinguish between the older type of political historians who tended to be quite suspicious of the social sciences and the "total" historians who found much that was useful in social-scientific generalizations. It was not so long ago, after all, that the "total" historians were the heretics and the founders of journals promoting new viewpoints. Nevertheless, for the problems under consideration in this paper, such differences are less important than the emphasis that members of both groups have placed on the autonomy of history or their shared objections to many social-scientific techniques.

However willing traditional historians might have been to turn to the social sciences for insights into human behavior, the majority recoiled from the analytical methods of the social sciences. The mathematical modelling and the preoccupation with measurement that were flowering in these disciplines were widely viewed as antihistorical, sterile, and a threat to the most intrinsic qualities of history—its literary art, its personal voice, and its concern for the countless subtle qualities that are involved in the notion of individuality. Most traditional historians valued literary art not only for its esthetic qualities but as an essential ingredient in conveying the experience of the past. They shared Theodore Roosevelt's belief that "unless he writes vividly" the historian "cannot write truthfully; for no amount of dull, painstaking detail will sum up as the whole truth unless the genius is there to paint the truth." Commitment to literary art did not rule out the exploitation of quantitative evidence, and the *Harvard Guide* has an extended discussion of this category of evidence and its uses. But quantitative evidence was generally considered of ancillary importance. Arthur Schlesinger, Jr. probably expressed the predominant view among traditional historians when he said that "almost all important questions

⁶ ELTON (1967), pp. 11, 24–26, 28, 29. For thoughtful assessments of the role of the social sciences in traditional history see HOFSTADTER (1956), HANDLIN *et al.* (1954), HIGHAM (1970, esp. ch. 1), and HANDLIN (1979).

are important precisely because they are *not* susceptible to quantitative answers".⁷

The reluctance of traditional historians to embrace statistical evidence was due partly, I believe, to the degree to which they had come to rely on the legal model for the evaluation of evidence. As Hans Zeisel has pointed out, courts have had a "distrust of statistical evidence," especially when offered as "proof of individual, specific events." Judicial distrust is due partly to fear that samples may be poor reflections of the universe they purport to represent and partly to a belief that statistical evidence was a form of hearsay evidence whose accuracy could not be tested easily by cross-examination. As a consequence legal doctrine, Zeisel noted, often allows the "testimony of selected witnesses who are far from constituting a representative sample but will refuse admittance of a survey based on a representative sample." (ZEISEL, 1968, p. 247.) The parallel with views held by traditional historians is striking and this suspicion of modern statistical procedures, as we shall see, constitutes one of the principal objections by the new brand of "scientific" historians to traditional historical methodology.

The model worked out in the 1930s, 1940s, and 1950s, and exemplified by the historians I have named, continues to be the model for traditional historians today.

The new brand of "scientific" history

The time has arrived for the introduction of the deuteragonist. The new brand of "scientific" history, which I will call "cliometrics," entered the historical lists during the 1950s. Although cliometricians are sometimes referred to as a "school," the term is somewhat misleading since cliometrics encompasses many different subjects, viewpoints, and methodologies. The common characteristic of cliometricians is that they apply the quantitative methods and behavioral models of the social sciences to the study of history. The cliometric approach was first given systematic development in economic history, but like a contagion it rapidly spread to such diverse fields as

⁷ HOFSTADTER (1968b, pp. 11-13), HANDLIN *et al.* (1954, pp. 22-30, 44-49), ROOSEVELT (1918, p. 486), SCHLESINGER, Jr. (1962, p. 770). Roosevelt's presidential address to the AHA remains one of the most powerful and insightful statements of the role of literary arts in the writing of history. It is obligatory reading for all who aspire to master the craft, whether they view themselves as scientific or traditional historians.

population and family history, urban history, parliamentary history, electoral history, and ethnic history.⁸

Cliometricians want the study of history to be based on explicit models of human behavior. They believe that historians do not really have a choice of using or not using behavioral models since all attempts to explain historical behavior—to relate the elemental facts of history to each other—whether called “Ideengeschichte,” “historical imagination,” or “behavioral modeling,” involve some sort of model. The real choice is whether these models will be implicit, vague, incomplete, and internally inconsistent, as cliometricians contend is frequently the case in traditional historical research, or whether the models will be explicit, with all the relevant assumptions clearly stated, and formulated in such a manner as to be subject to rigorous empirical verification.⁹ The approach sometimes leads cliometricians to represent historical behavior by mathematical equations, and then to seek evidence, usually quantitative, capable of verifying the applicability of these equations or of contradicting them. The behavior that cliometricians have dealt with so far has generally been represented by single equations or by simple simultaneous-equation models with relatively few variables. These equations are usually linear in form or involve linear or other low-order approximations.¹⁰

⁸ For efforts to move in this direction prior to World War II see HECKSHER (1929), BRINTON (1930), CLAPHAM (1931), and ROSTOW (1948). Discussions of the characteristics of cliometrics include ANDREANO (1970), AYDELOTTE (1966, 1971, 1973), AYDELOTTE *et al.* (1972), BENSON (1972); CLUBB and BOGUE (1977), CONRAD and MEYER (1964) CLUBB (1975), CLUBB and ALLEN (1967), DAVIS (1968, 1971), DOLLAR and JENSEN (1971), FOGEL (1970, 1975), FLOUD (1973), FURET (1971), FURET and LADURIE (1970), GALLMAN (1971), GOULD (1964), HABAKKUK (1971), HAYS (1968), J. HUGHES (1966), JENSEN (1974); KAHK and KOVALCHENKO (1974), KOUSSER (1976, 1977, 1979), LAMPARD (1975), LANDES and TILLY (1971), LORWIN and PRICE (1972), NORTH (1963, 1968); McCLELLAND (1975), MCCLOSKEY (1978), REDLICH (1965), ROTHSCHILD *et al.* (1970), ROWNEY and GRAHAM (1969), SAVETH (1964), SPRAGUE (1978), SWIERENGA (1970, 1974); TEMIN (1966), TILLY (1978), VINOVSKIS (1978), WRIGLEY (1972).

⁹ See, for example, BENSON (1972), BOGUE (1968), CLUBB and ALLEN (1977), DAVIS (1968), FOGEL (1967), SMITH (1979), WILLIAMSON (1974).

¹⁰ On the role of mathematics in cliometric research see ELSTER (1978), FOGEL (1970 and 1975), KAHK and KOVALCHENKO (1974), KOUSSER (1976 and 1979), LAMPARD (1975), McCLELLAND (1975), MILOV and KHOVOSTOVA (1973), KOVALCHENKO *et al.* (1977), SPRAGUE (1978). On the relationship between mathematical models and the empirical assessment of counterfactual-conditional statements, see BRAYBROOKE (1977), ELSTER (1978), ENGERMAN (1979), FOGEL (1970), GOULD (1969), McCLELLAND (1975), MURPHY (1969), SIMON and RESCHER (1966).

Such mathematics might be thought to be too simple to be useful as a characterization of complex human behavior. Nevertheless, actual practice has shown that this simple mathematics is often a powerful instrument in advancing knowledge of the past. First, by making the assumed behavioral relationships explicit, these models lay the basis for a considered discussion of the circumstances under which linear or other lower-order approximations or more complex relationships are adequate or inadequate. Quite often the narratives of traditional historians, when dealing with relationships between variables, implicitly assume the most simple of all functions—strict proportionality between the variables. It has been shown that when this severe restriction is relaxed, and a more realistic functional relationship is introduced, the interpretations of some historical events are greatly altered. Much of the work of the cliometricians has been directed to spelling out and formalizing the models implicit in traditional historical narratives and to a consideration of the empirical validity of those models.

Second, the mathematical characterization of historical behavior has helped to identify the critical parameters in historical narratives. Because of incompleteness of data, historians frequently have widely different beliefs about the values of the parameters that implicitly or explicitly enter into their analyses. Translating such arguments into mathematical form makes it possible to engage in "sensitivity analysis,"—that is, to examine the sensitivity of the conclusions of an argument to alternative estimates of particular parameters. This procedure has eliminated many unnecessary wrangles by demonstrating that the absence of exact information on particular points is at times inconsequential. For quite often a measurement which is logically necessary for a given analysis may be such that any plausible number, even though it may deviate greatly from reality, is permissible and serves to close the logical system on which the analysis is based. Albert Fishlow, for example, employed this device in his reconstruction of the U.S. pattern of interregional trade before the Civil War by guessing at the share of southern imports that were re-exported and then demonstrating that no plausible error in his guess could alter his results by more than a few percentage points.¹¹ While such techniques do not eliminate all errors or banish all needless wrangles, they reduce them by providing criteria that facilitate the identification of error and the resolution of issues.

¹¹ FISHLOW (1964, Appendix). Of course, sensitivity analysis often reveals that small changes in estimated parameters may radically alter an interpretation. Cf. POPE (1975).

It is not analysis but description that occupies most of the time of cliometricians. In this respect, cliometricians conform to Ranke's admonition that historians should devote themselves to the task of determining what actually happened. Just as the nineteenth- and early twentieth-century followers of Ranke scoured the public archives for diplomatic and ministerial documents that would reveal what actually happened in government policy, so cliometricians have been scouring archives anew, this time searching for quantitative evidence bearing on what actually happened in social behavior.

And so we arrive at the crux of the difference between traditional history and cliometrics. Many traditional historians tend to be highly focused on specific individuals, on particular institutions, on particular ideas, and on nonrepetitive occurrences; those who attempt to explain collective phenomena generally make only limited use of explicit behavioral models and usually rely principally on literary evidence. Cliometricians tend to be highly focused on collections of individuals, on categories of institutions, and on repetitive occurrences; their explanations often involve explicit behavioral models and they rely heavily on quantitative evidence. A traditional historian, for example, might want to explain why John Keats died at the time, in the place, and under the particular circumstances that he did. But to a social-scientific historian attempting to explain the course of mortality among the English, the particular circumstances of Keats's death might be less interesting than those circumstances that contribute to an understanding of why deaths due to tuberculosis were so frequent during the first half of the nineteenth century.¹² Of course, these approaches are neither mutually exclusive nor in any sense antagonistic, although partisans of the two approaches often behave as if they were.

Some scholars treat quantification as *the* characteristic that identifies cliometricians. Quantification is more commonly encountered in cliometric work than the explicit mathematical modeling of behavior, but it is not a universal characteristic of such work. The term "cliometrician" embraces scholars who, although they rarely use numbers or mathematical notation, nevertheless base their research on explicit social science models. For reasons already suggested, and more fully discussed in the next section

¹² The scope of cliometric research in historical demography is indicated in DRAKE (1969), EASTERLIN (1977), FLINN (1970), FLINN *et al.* (1977), GLASS and EVERSLY (1965), IMHOFF (1978a and 1978b), LOCKRIDGE (1977), LEE (1977), McKEOWN (1976), SMITH (1977), TILLY (1978), VANN (1969), SCHOFIELD and WRIGLEY (1980), VINOVSKIS (1978).

of this paper, no single characteristic can be used to distinguish between traditional and "scientific" historians, although a scholar's attitude toward the autonomy of history may go further in that direction than any other particular characteristic.

Methods of authenticating evidence also serve to distinguish the two groups. The methods that traditional historians have developed for authenticating evidence were geared more to specific events involving specific individuals than to repetitive events involving large groups of individuals. Of what use is the criterion that an historian must seek two corroborating opinions, if the point at issue is whether the standard of living of the English working class declined during the Industrial Revolution? There were scores of different opinions on this question and even a fairly unresourceful historian would have no difficulty in discovering two or even several witnesses who shared a particular view. But such a limited degree of accord could also be established for directly opposed views.¹³

Indeed, the very concept of decline in the standard of living is much different for a group than for an individual. Even in the worst depressions, when the economic circumstances of most individuals are deteriorating, there are some individuals whose economic circumstances are improving. Methods of analysis that are appropriate for determining whether George Washington's income declined during the post-Revolutionary years will not do for the determination of whether or not the income of American slaveowners as a class declined.

Simple transposition of techniques that work quite well for the analysis of individual behavior may do more to distort than to clarify collective behavior. An individual who had so politically split a personality that he simultaneously embraced the policies of a revolutionary party and their most ardent opponents would be classified as psychotic. Yet such split behavior is normal in the case of nations, churches, classes, and other substantial social, political, and economic formations. Whatever the qualities that give individuals a group identity, one never encounters such uniformity in their positions, attitudes, and responses that they can be treated as having an identical personality. Explaining the behavior of a parliament, which has a large number of members, thus frequently poses problems which are quite different from the explanation of the behavior of an absolute monarch, of a prime minister of a democratic government, or of a few of the leaders in a parliament.

¹³ See A. J. TAYLOR (1956) for a survey and assessment of this debate to 1975.

This point is recognized by both traditional and "scientific" historians, and both groups have sought to come to grips with the problems of studying collective behavior in their own ways. In dealing with parliaments, traditional historians have tended to rely on the opinions of individuals who were at the center of parliamentary struggles, and so were in a position to know what was going on, or who, although just observers, were of such keen mind that they were likely to have grasped the essence of the situations. Scientific historians have tended to concentrate on the analysis of roll calls and on quantifiable characteristics of legislators or their constituencies. They have sought statistical methods which are capable of squeezing from such evidence information on the existence of blocs in parliaments and parties, the intensity of adherence to various positions, the underlying factors which give unity to (or threaten to destroy) coalitions, and the mind set of particular legislators or categories of legislators.¹⁴

The cliometric approach would be of some interest even if it merely confirmed what had already been discovered by traditional historiographic methods. In virtually every field to which it has been applied, however, the cliometric approach has not only yielded substantive findings that are strikingly different from the findings of the older research but has also called attention to important processes that previously had escaped notice.

Take the study of family and household structure that was launched by Peter Laslett and his colleagues in the Cambridge Group for the History of Population and Social Structure in 1965, and which has since stimulated similar research programs in the Scandinavian countries, the Low countries, Germany, and North America (LASLETT, 1972, 1977). The first of many surprises to emerge from this work was the discovery of an extraordinary degree of geographic mobility among residents of preindustrial English villages, which led, on average, to a turnover of more than 50 percent in the population of such villages per decade. More far reaching was the discovery that the so-called nuclear or simple household form of family organization (kin limited to parents and children) is not the recent product of highly industrialized society, as previous scholars had suggested, but has been the predominant form of household organization in northwestern Europe and the United States for at least three hundred years.

Laslett and his associates believe that they have discovered a "Western" family pattern that lasted for several centuries, and was quite distinct from

¹⁴ The scope and methods of clicmetric research in political history are indicated in AYDELOTTE (1977), BOGUE (1978), SILBY *et al.* (1978), TILLY (1975).

the complex family and household pattern (presence of kin other than parents and children) that until recent times prevailed in some parts of Eastern Europe. The "Western" pattern was distinguished by the nuclear household form, by a relatively late age of marriage (generally in the late twenties or early thirties for women), by a relatively high proportion (20 to 25 %) of wives older than husbands, and by the presence, in a relatively high proportion of the ordinary households, of non-kin servants who lived and worked in the household from the time of puberty until the time of marriage. By contrast, East-European households, particularly in Russia, appear to have been extended and multigenerational, marriage was at a relatively young age (generally in the late teens or early twenties for women), and there was an almost complete absence of non-kin servants and boarders from ordinary households. These and other findings, although challenged by a number of scholars, have stimulated far-ranging discussions of the possible effects on the development of personality and of such important cultural questions as evolving attitudes toward childhood and other domestic ideals.¹⁵

Or take the debate over the economics of the U.S. slave system, which began in the mid-1950s, and has continued with great intensity down to the present time. Until the mid-1950s it was widely believed that the slave plantations were unprofitable and inefficient enterprises that were kept in operation by a class prepared to sacrifice its private economic interest, and to endure economic stagnation for the South, in order to maintain its political and cultural hegemony.

The first critical blow to this thesis was delivered by the now famous 1958 essay of Alfred H. Conrad and John R. Meyer. Marshalling the limited quantitative evidence available at that time on inputs, outputs, and prices, and using a standard capital model to estimate the internal rate of return on an investment in slaves, they concluded that rates of return on this form of capital compared favorably with rates on alternative investments. This finding set off a passionate debate that took more than a decade to resolve. Eventually refinements in the original capital model and a considerable expansion of the data base, including the Parker-Gallman sample of over 5,000 southern farms and large samples of slave prices and hire rates from probate and other records, placed the rate of return in the 6 to

¹⁵ Other cliometric contributions to the history of the family are discussed in HARAVEN (1977), HARAVEN and VINOVSKIS (1978), VINOVSKIS (1977), WACHTER *et al.* (1978), WRIGLEY (1970).

10 percent range, thus demonstrating the allocative efficiency of the slave economy. The second blow to the traditional thesis was delivered by Richard A. Easterlin's development of regional income estimates extending back to 1840. Both these estimates and the subsequent refinements in them showed that the South experienced a high rate of growth in per capita income between 1840 and 1860, thus contradicting the view that the antebellum South was economically stagnant.¹⁶

These twin blows to the traditional interpretation shifted cliometric attention to the question of the relative technical efficiency of slave and free agriculture in input utilization. While it is premature to declare this question settled, a consensus has emerged on one critical point. It is now clear that the previous view that slave agriculture was less efficient than free agriculture is incorrect. What remains to be resolved is the exact margin of the advantage enjoyed by slave plantations, and the explanation for this margin.¹⁷

The new discoveries and challenges to old interpretations caught the imagination of younger scholars trained in both history and the social sciences. The cliometric approach developed most rapidly in economic history and has been the predominant form of research in this field, at least in the United States, for over a decade. The majority of the articles published in the main economic history journals of the United States are now quite mathematical and cliometricians predominate in the leadership of the Economic History Association. The rapidity of this transformation is due partly to the encouragement given to the new mode of research by the older, more traditional economic historians who generally welcomed and encouraged the experiments of their younger colleagues.¹⁸ The large library of economic models upon which cliometricians could draw and the relative ease of applying these models to the issues of economic history also contributed to the pace.

¹⁶ The debate over the profitability of slavery is reviewed in FOGEL and ENGERMAN (1974, vol. 1, ch. 3, and vol. 2, pp. 54-87). The debate on antebellum southern economic growth is reviewed in GALLMAN (1979). Cf. GENOVESE and FOX-GENOVESE (1979), WOODMAN (1972), WRIGHT (1978).

¹⁷ Contributions to the debate on the efficiency of slave agriculture include DAVID *et al.* (1976), DAVID and TEMIN (1979), FOGEL and ENGERMAN (1974, 1977, 1980), WRIGHT (1979).

¹⁸ Assessments of the development of the cliometric approach in economic history include ANDREANO (1970), COCHRAN (1969), COLEMAN (1977), ENGERMAN (1977), HABAKKUK (1971), HARTWELL (1971), J. HUGHES (1971), LÉVY-LEBOYER (1969), MATHIAS (1971), McCLOSKEY (1971 and 1978), NORTH (1977 and 1978), PARKER (1972).

Cliometricians working in political, social, and intellectual history faced a more difficult challenge. The problems they sought to master were inherently more difficult than those in economic history and the library of social science models on which they could draw was more skimpy. They also encountered far more resistance to their efforts among the traditional historians in their fields and the editors of the mainline U.S. history journals showed little interest in their papers. As a consequence, a number of new journals which catered to the new mode of research came into being and have flourished: among them, *Historical Methods*, the *Journal of Social History*, the *Journal of Interdisciplinary History*, and the *Journal of Family History*. The pages of traditional journals have recently become somewhat more open to quantitative studies and most of the major departments of history in the United States now include on their faculties one or more scholars who work in the new mode.

For many cliometricians the progress of integration has seemed much too slow.¹⁹ A recent survey of 83 U.S. history departments by J. Morgan Kousser revealed that 64 percent offered a course in statistical methods to graduate students, but in nearly all cases the level of the course was quite elementary. Only 10 percent of the departments offered courses in social-scientific theory that could teach students the art of applying behavioral models to the study of history even at an elementary level. While such courses might, as Kousser points out, overcome the "math anxiety" that now haunts historians, they do not prepare graduate students to conduct serious research in the cliometric mode.

Many cliometricians want sweeping changes in both the graduate and undergraduate history curriculums. But with a few exceptions (as at Pittsburgh, Caltech, and Carnegie-Mellon) sweeping changes have been blocked by traditional historians, some of whom believe that anything more than a superficial acquaintance with the social sciences is unnecessary and may be dangerous. As Lawrence Stone recently put it, an historian should come to the social sciences only "as a seeker after a specific idea or piece of information" and that more intense training in social science methods would "make it impossible" for history graduate students "to obtain that broad historical knowledge and wisdom and that familiarity with the handling of sources which have hitherto been regarded as essential prerequisites for the professional historian" (STONE, 1977b, pp. 18, 37).

¹⁹ The discussion in this and the next two paragraphs is based largely on KOUSSER (1980).

Impatient with the slow progress of curriculum reform within their departments, U.S. cliometrists have sought to bypass traditional channels by setting up various intensive summer training programs to which they could send their students. One of these has now been going on for eleven summers at the University of Michigan and another for eight summers at the Newberry Library in Chicago. The climax of these efforts to accelerate the swing toward cliometrics was the establishment in 1975 of the Social Science History Association (SSHA), which now has about 700 members and whose annual conferences bring together several hundred scholars. SSHA also publishes its own journal (*Social Science History*) and attempts to coordinate its activities with those of like-minded scholars abroad.

The existence of two well-defined modes of historical research

And so what we have in history today are two well-defined, competing research modes—paradigms, for those who prefer Kuhnian language—which for convenience I have called “traditional history” and “scientific history.” Each mode has a set of traditions that defines criteria of excellence, and a set of problems considered the appropriate subjects of research; each has its own methodology and set of intellectual concepts (cf. KUHN, 1970). I now want to define the two modes of research by counterposing the characteristics that distinguish them more systematically than I have so far. The principal differences may be summarized under six headings: subject matter, preferred types of evidence, standards of proof, the role of controversy, attitudes toward collaboration, and communication with the history-reading public.

In the discussion that follows, as well as in that which has already taken place, my effort to describe the attitudes and approaches of traditional and “scientific” historians must be viewed in a statistical light. While I may characterize the approaches of each group on one or another question by what I perceive to be the central tendency of its practitioners, it would be quite erroneous to conceive of all of the practitioners as being located at that point. Even though most cliometrists probably agree that statistical methods are a prime instrument in historical analysis, their assessments range from “indispensable on most issues” to “useful on some issues.” Among traditional historians today the range of views on this question extends from “often useful on many issues” to “rarely useful and often harmful,” with most probably agreeing that when quantitative methods

are used with care and restraint, they are helpful instruments of secondary importance.

Because of the wide range within each group, there are frequent overlaps of view from question to question and a considerable number of scholars from each group fall into these areas of overlap. Those with a mind to do so could no doubt develop fine arguments over whether scholars such as Richard Hofstader, Oscar Handlin, Lewis Namier, and Fernand Braudel are best viewed as traditional historians who make considerable use of social science or as "scientific" historians who make considerable use of traditional approaches. It would be quite unproductive to quarrel over where to put those who might easily fit into more than one category, and a complete waste of time to dispute a scholar's own view of this matter.

Subject matter. This may be the most fundamental point of difference. "Scientific" historians tend to focus on collectivities of people and recurring events, while traditional historians tend to focus on particular individuals and particular events. I do not mean to suggest either that "scientific" historians do not study particular events or that traditional historians do not study social and political movements. But when "scientific" historians study the stock market crash in 1929, the decision of the British Parliament to end slavery in its colonies, or the downfall of Louis XVI, they proceed on the assumption that these particular events were the outcome of processes that were governed by functional relationships containing both systematic and stochastic terms. Their manner of approach to particular events is quite similar to that being employed by the National Transportation Safety Board in its investigation of another particular historical event, the crash of a DC-10 at O'Hare Airport in Chicago. The agency seeks to determine, on the basis of a knowledge of laws of aerodynamics, properties of materials, etc., and as precise a knowledge of the facts associated with the event as possible, the most likely explanation for the crash (cf. JOYNT and RESCHER, 1961).

Of course, traditional historians are also concerned with general forces and they recognize that these have an impact on human behavior. But they would not accept as an explanation of Louis Napoleon's decision to go to war with Prussia a statement of the form: "that under such and such circumstances monarchs are likely to go to war." Even if such a general statement correctly recorded that 7 out of 10 monarchs would declare war under similar circumstances, it would not explain why *Louis Napoleon* went to war. While a traditional historian would deal with the underlying

forces that set the stage for the Franco-Prussian war, he would want to know why Louis Napoleon was more influenced by the hawks than the doves and whether he was fooled by intrigues or convinced by a weighty set of arguments. He would also want to know how important the influence of Napoleon's strong-willed wife was and whether his gout and other ailments had anything to do with his acquiescence.

Another way of making the same point is to say that traditional historians often concentrate on problems in which the influence of the stochastic terms are predominant, while "scientific" historians often concentrate on problems in which the systematic terms are predominant. What is background for one group is the central concern of the other. Some scholars might be inclined to argue over whether the systematic or the stochastic elements ought to be the principal focus of historical research. The answer will surely vary from case to case and will in part depend on which aspect has been most fully explored by previous scholars.

This difference in orientation helps to explain the impatience that cliometric and traditional historians have with each other's research agendas. Few traditional historians will deny that some information on vital rates is useful, but they are bored by cliometric preoccupation with the temporal patterns of fertility and mortality rates, and by the endless debates over the timing of changes in the two series. They doubt that much useful information will come from international comparisons of data on household structure, from correlating changes in the land-to-labor ratio with the rise and decline of feudalism, from tracing collective violence, or from the analysis of parliamentary roll calls. Issues that are at the very top of some cliometric research agendas, such as the decade-long debate over the effect of the frontier on the nature of U.S. technology, do not appear even to have penetrated the consciousness of traditional historians. Nor is there much evidence that issues of such high current concern to traditional historians as pre-industrial *mentalité*, the ideology of the abolitionists, or the evolution of religious thought are as yet of professional interest to many cliometricalians.

Of course, there are questions that have attracted the attention of both traditional and "scientific" historians and in which both groups of scholars have paid more attention to the underlying social, economic, and political forces than to contingent factors or particular personalities. This has certainly been the case when both types of historians have sought to account for the origins of the New South, the changing nature of love and the family in Great Britain, or the decline of feudalism in Western Europe.

In such instances it is the other considerations that I listed, particularly preferred types of evidence and standards of proof, that distinguish the two modes of investigation.

Preferred types of evidence. Both traditional and "scientific" historians must work with the surviving evidence, usually of a documentary nature, that posterity has bequeathed, but there are differences in the types of evidence that are most likely to satisfy them. Traditional historians have exhibited a strong preference for literary evidence while "scientific" historians lean strongly toward quantitative evidence. To answer the question, how strong were the bonds of affection that tied members of slave families to each other, a traditional historian might call witnesses such as Harriet Beecher Stowe, whose Uncle Tom's Cabin gave testimony to the warmth and affection that suffused slave families. Another traditional historian could counter that equally qualified witnesses, such as Fanny Kemble, testified that the slave system was so ferocious, so destructive in its effects, that it stripped slave families of all "tenderness" and "spiritual grace" (HANDLIN, 1975, p. 13; KEMBLE, 1961, p. 95).

In such cases of conflicting testimony (normal with respect to large groups) most cliometricians prefer information on actual behavior to opinions about that behavior, and they would seek data bearing on the entire distribution of family responses to the slave system. Such information is needed to know which witnesses were describing typical behavior and which were describing aberrant behavior. Cliometricians have thus sought information on the structure of slave households; on the distribution of the length of slave marriages; on the percentage of disrupted slave families that reunited when they were free to do so; and on the distribution of the characteristics of the households in which slaves were raised, as reported in the thousands of narratives of ex-slaves (CRAWFORD, 1980; HIGMAN, 1976, 1978; SUTCH, 1975; STECKEL, 1977; TRUSSELL and STECKEL, 1978; cf. CRATON, 1968; GUTMAN, 1976). The traditional historian would discount such evidence because of doubts over the representativeness of the samples; he would wonder whether the memory of aged ex-slaves could be trusted, and whether their replies were affected by the race of the interviewer. The "scientific" historian would attempt to devise means of estimating the magnitudes of such possible biases.

Scholars who do not like the outcome of a head count often assert that the count does not count because the source is rife with error and hence biased (in the statistical sense). Frequently this assertion is put forward

on the basis of the most tenuous evidence, or no evidence at all—sometimes just on the basis of *ad hominem* arguments; sometimes on *a priori* contentions that it is reasonable to assume the existence of biases. Attacks on bodies of quantitative evidence should not necessarily be accepted on face value but require as careful an evaluation as any other allegation. Historians are not always as demanding of those who call for the dismissal of a large body of quantitative evidence, on the ground that it might be defective in one or another respect, as they are of those who offer the evidence.

It is not always appreciated that even proven defects in a given body of evidence do not necessarily deprive it of usefulness. Defects may invalidate the source on one issue but not on another, for one purpose but not for another. If, for example, a given register of vital events is known to be incomplete in its coverage of births because of the omission of those who died soon after birth, it could still be used to establish a lower bound on the birth rate or an upper bound on the average length of the birth interval, and it could still yield an unbiased estimate of the rate of natural increase. Such limited information may be sufficient for many historical issues. There is also the question of the magnitude of the biases. Sometimes it is possible to demonstrate that the defects in a given body of data are too small to have a significant effect on the analytical issues that will be addressed to it. So, before dismissing a large body of evidence it is important to determine the nature of the errors that afflict it. It makes a considerable difference whether the error is random or systematic. A body of quantitative evidence will yield unbiased estimates of the parameters at issue, although it is rife with error, when the error is randomly distributed. Even when the data are afflicted by systematic bias, it may be possible to devise techniques that will yield acceptable estimates of desired parameters. James Trussell and Richard Steckel have shown that it is possible to estimate the age of slave mothers at the birth of their first child with a negligible degree of error, from data known to suffer from substantial underrecording of births, by making use of a statistical technique called the “singulate mean.”

The worst of all errors is to assume that either literary evidence by itself or quantitative evidence by itself is sufficient, when they are not. Such self sufficiency cannot be achieved when the object of study is a broad social movement. Moses Finley is surely right when he says that “all the possible statistics about age of marriage, size of family, rate of illegitimacy, will not add up to a history of the family.” That history must deal fully with a series of issues about the quality of family life, such as those that

Stone has addressed: the changing roles of husbands, wives, and other kin and of relationships between them; their changing attitudes toward each other; and the effects of family attitudes and roles, not only on the culture of families and the fate of its individual members, but their broader repercussions for society, economy, and the state. It is also true, however, as Finley has stressed on other occasions, that even the most subtle and imaginative discussions of these issues will collapse if they are not based on a rock of evidence as to what is typical and what is aberrant. Stone's views on the array of issues that he tackles (as Keith Thomas, Christopher Hill, and Alan Macfarlene, among others, have emphasized) are time and again anchored on assumptions as to what was typical and what was not, on assumed trends in household structure and in demographic and economic variables, and on the assumed interrelationships between these variables and such subtle matters as sentiment, commitment, obligation, and emotional texture. Can it be denied that a satisfactory history of the family must have both qualitative and quantitative aspects, and that neglect of either may lead the historian astray? (FINLEY, 1977, p. 139.)

Standards of proof and verification. The traditional historian's model for proving his case or disproving an opponent's case is the legal model. (HANDLIN *et al.*, 1954; cf. BERMAN, 1968.) He seeks to employ witnesses of high moral character and ability, and attempts to show that they were in a position to know what happened. He disputes his opponent's witnesses by impugning their character, their objectivity, and their capacity to know. Similar standards are applied to documentary evidence. Traditional historians also follow the legal tradition of reasoning by analogy, of attempting to show that one situation is quite analogous to another. Eugene D. Genovese in *Roll Jordan Roll*, for example, finds an analogy between slave craftsmen, and the artisans who sparked the radical movements in eighteenth century England, France, and the American North. This analogy leads to the conclusion that slave craftsmen "provided the firmest social basis for a radical political leadership."²⁰ Traditional historians also follow legal precedent by the invocation of authority. When Engerman and I computed an index of total factor productivity to test

²⁰ GENOVESE, 1974, p. 394. According to DROYSSEN (1893, p. 97) reasoning by analogy was intrinsic to all historical understanding. "By being in motion, as ourselves are within, the world without permits us to understand it under the analogy of that which is going on in ourselves".

opinions of authorities who claimed that slave labor was inefficient, some critics replied that the computation must have been in error, because it was in disagreement with the opinions of established authorities.²¹

The cliometrician's model for proving his case or disproving an opponent's case is the empirical-scientific model. The strategy is to make explicit the implicit empirical assumptions on which many historical arguments rest and then to search for evidence, usually quantitative, capable of confirming or disconfirming the assumptions. On the question of the profitability of slavery, for example, Conrad and Meyer argued that testimony from planter diaries and similar sources were contradictory and inadequate. They set out to resolve the issue by obtaining representative samples of data on the output and prices of cotton; on the quantities and prices of the land, equipment, and other capital used in production of output; on slave maintenance costs; and on slave death rates. The long debate on their computation of the rate of return turned on such issues as the representativeness of the samples, the completeness of the coverage of outputs and costs, the adequacy of the mortality schedules, and the sensitivity of the computations to alternative methods of setting up the equation to calculate the rate of return.²²

The differences in the types of evidence on which proof rests affect the process of scholarly verification. When proof depends on the testimony of witnesses, the footnote is the critical element of documentation. It directs the skeptical reader to the location of the testimony and allows him to form his own judgment of the relevance of the testimony, its reliability, and

²¹ See, for example, HASKELL (1975a), SCHEIBER (1975), McCLELLAND (1978). McClelland argued that when our result clashed with established authority, Engerman and I should have realized that we were in error. He suggested that the source of the error was the use of a measure of efficiency derived from a Cobb-Douglas production function. But as we have shown (FOGEL and ENGERMAN, 1980), our result is independent of the form of the production function. The conclusion that slave farms employing the gang system were more efficient than small free farms follows merely from an examination of the location of the production points in production space. This conclusion holds with even greater force when we use the alternative measure of efficiency proposed by WRIGHT (1979).

²² CONRAD and MEYER (1964, pp. 43-114), FOGEL and ENGERMAN (1974, II, pp. 54-87). Conrad and Meyer were not the first to argue that slavery was a profitable investment for slaveholders. Gray and Stampf, among others, had strongly argued that view. But while the evidence they marshalled supported their case, it was inadequate to resolve the issue. See WOODMAN (1963) for an assessment of the status of the debate after the appearance of the Conrad and Meyer paper, but early in the stream of research aimed at evaluating the validity of their approach and findings.

the judgement that was exercised by the scholar in interpreting the testimony.

In cliometric research, where large bodies of data are the basis of proof, the role of the footnote is greatly diminished, although it is still used to inform the reader about the kinds of materials used and for other ancillary purposes. The critical information is usually conveyed in tables and charts or in equations and it is impossible to report in footnotes the thousands or tens of thousands of observations from which these were constructed. Even the procedures followed in the analysis of the data cannot usually be reported in detail in the same study since it frequently involves hundreds or thousands of operations. Consequently, the usual practice of cliometrists is to make their data available to other scholars by reproducing their computer tapes upon request or by putting them on deposit with the Inter-University Consortium for Political Research at the University of Michigan, which maintains an international lending library of computer tapes. The procedures employed in analyzing the data on the tapes are often reported in separate technical papers (and sometimes published long before the publication of the substantive work) or else are described in mimeographed papers and code books, or in worksheets, which are available from the investigators on request.

These procedures create no great problem for cliometrists who are used to requesting computer tapes, format statements, code books, and worksheets when they desire to replicate the analysis of a colleague—sometimes in order to build upon it, sometimes to dispute it. But traditional historians are often appalled by the effort that is required to verify cliometric research. This is what Lawrence Stone said when he contemplated the verification of the graphs in *The Rebellious Century, 1830–1930* by Charles, Louise, and Richard Tilly (STONE, 1977b, p. 31):

In order to discover the sources and methods that lie behind graphs 5 to 8, on acts of collective violence in France over a century—the collection, coding and analysis of which took countless man-hours of many researchers over almost a decade—the reader is asked to track down descriptions of the methodology spread over no fewer than six different articles (p. 314). Few readers will have the tenacity or curiosity to pursue the subject that far. The great majority will inevitably take the graphs at their face value, without probing any deeper. The major findings of the work stand or fall on the reliability of these graphs, and yet within the book itself there is no provision

made for discovering how they were compiled, while the multivariate analyses used for explaining the ups and downs on the graphs are likely to baffle all but the most sophisticated of cliometricians. This is a book which lacks most of the basic scholarly apparatus but which apparently conforms to the best standards of scholarship of which cliometric history is capable. It is the product of a decade of massive research, and yet the reader is left in a state of helpless uneasiness both about the reliability of the data and about the validity of the explanations put forward. It therefore poses in its starker form the problem of verification in cliometric history.

The role of controversy. While traditional history has had its share of controversies, controversy is not normally the mark of a successful study. Success normally turns on how widely and how well a work is received. Although there are notable exceptions, strong attacks, especially if they come from distinguished colleagues, tend to undermine the credibility of a work even if the attack is empirically unwarranted. The traditional historian often comes before his colleagues and the history-reading public as an expert witness who has carefully examined all of the issues and his book or paper constitutes his expert testimony—his deposition, so to speak. Thus distinguished traditional historians sometimes depart from the monographic pattern of documenting each statement in their study by footnotes, presenting only a bibliography of the sources they examined. An attack on the credibility of the historian *qua* witness, and many of these attacks are *ad hominem*, has the same force as an attack on a witness in court. It diminishes his testimony.

Controversy is rife among “scientific” historians and many traditional historians have interpreted the sharp disagreements as evidence of the failure of social science methodology, and particularly of quantitative methods, in history. This view reflects a confusion between artistic and scientific processes, and brings us back to the significance of the artistic element in traditional history. A painting, a concerto, a novel, and many traditional histories can be the perfect creation of a single individual during a relatively brief period of intense activity. As with artistic works generally, traditional histories normally have a highly personal quality. Scientific creations, however, are usually protracted over long periods, approach perfection quite gradually, and often involve the efforts of a large number of investigators. Such controversies as those over the explanation for the demographic transition in Great Britain, the economics of U.S. slavery,

the transition from serfdom to bourgeois agriculture in Russia, the structure of households in Northwestern Europe, social mobility in cities, entrepreneurial failure in Victorian Britain, the decline of fertility in the United States, and the social saving of railroads have demonstrated the great complexity of the analytical issues, the large amounts of data that must be retrieved to resolve them, and the many pitfalls that may be encountered in the analysis of these data. Such problems are resolved through collective effort, one aspect of which is the intense debate over the significance and validity of successive contributions.²³ One should not imagine that scientific controversies are necessarily free of invectives or necessarily lead to full agreement among all parties to it. A "new scientific truth," as Max Planck once remarked, does not necessarily "triumph by convincing its opponents" (cited by KUHN, 1970, p. 151). Perhaps more often than not, it does so by convincing the next generation of specialists who, because they are not so personally involved, can view the dispute with objectivity.

While traditional historians tend to accept or impeach an historical work on the totality of its interpretation, cliometricians tend to assess each estimating procedure and each result in a large work separately. On important questions several different cliometricians may attempt to replicate a given result. Frequently, the interactions between critics and investigators will go through several rounds, with each round refining more precisely the points at issue and calling for more exact computations or more detailed data. That is why the process of verification, which may also be a process of modification, is often so protracted. Cliometricians tend to place a high premium on findings that surprise them, but such findings are likely to undergo the most searching criticisms before a consensus on their validity is achieved.

²³ On the debate over the demographic transition in Great Britain, see DRAKE (1969), FLINN (1970), LEE (1977), McKEOWN (1976), SCHOFIELD and WRIGLEY (1980), RAZZELL (1974). On the debate over economics of U. S. slavery, see the sources cited in notes 16 and 17. On the debate over the structure of households in Europe, see LASLETT (1977), WACHTER *et al.* (1978) and the sources cited in these. On the debate over social mobility in the U.S., see ALCORN and KNIGHTS (1975), ENGERMAN (1975), SCHNORE (1975), THERNSTROM (1973 and 1975). On the debate over the transition in Russian agriculture, see FIELD (1969), METZER (1977), KOVALCHENKO (1967), and the sources cited in these. On the debate over British entrepreneurial failure see MCCLOSKEY (1971). On the debate over the decline of fertility in the U. S., see BOGUE (1976), EASTERLIN (1976, 1977); POTTER (1965), VINOVSKIS (1978). On the debate over the social saving of railroads see HAWKS (1970), O'BRIEN (1977), FOGEL (1979).

These observations suggest the different roles of controversy in traditional and "scientific" history, but they do not define the difference adequately and it is unlikely that any brief discussion can do so. Although many of the characteristics that distinguish the two research modes are revealed in their controversies, the manifestations are subtle and complex. One cannot, for example, adequately discuss the impersonal, sometimes arid quality of cliometric debates, which repel so many traditionalists, without setting forth a host of qualifications. It is necessary to call attention not only to such exceptions as the emotional storm over the economics of slavery, but also to deal with the connection between tone and substance. The impersonality of cliometric debates is related to the relatively limited range of the points at issue and to the high proportion of the points that can be resolved, at least in principle, by measurement and other scientific procedures. The well-defined research agendas that emerge from such controversies tend to be focused on technical points which, however remote they may seem to be from the deeply human issues that originally sparked the debates, eventually turn out to involve matters of considerable consequence.

Technical points are also involved in the debates of traditional historians—witness the disagreement between H. R. Trevor-Roper and Stone (1948, 1952) over the interpretation of forfeiture penalties in the bonds of indebted peers or Elton's work on the endorsements of state papers—but many other categories of issues are present as well. The ideological stance of a work, the quality of mind of its author, and stylistic merit, which loom so large in traditional disputes, seldom enter into cliometric history, just as they seldom enter into science proper. The last point raises still another complicated problem, the influence of external models on the styles of controversy. I have already alluded to the role of the legal model in traditional history, but in so far as the style of controversy is concerned, the influence of literary criticism may be more important. Cliometrists have been heavily influenced by the intellectual style of the physical sciences, although the influence of the rhetorical arts is also evident.

To emphasize the somewhat different roles of controversy in traditional and "scientific" history is not to imply that traditional history is bereft of novelty. Although novelty is sometimes introduced into traditional history through sharp attacks on reigning views (as in Beard's economic interpretation of the U.S. constitution, Elton's rehabilitation of Thomas Cromwell, or Namier's reinterpretation of the nature of British politics

during the late eighteenth century), it more often enters through the filling in of blanks in the story of history, and especially in opening up new lines of investigation. As Felix Gilbert has pointed out, "one of the great tasks and achievements of nineteenth-century historical scholarship was to establish the main features of the history of European nations from the ancient world to the eighteenth century and to place the story of their development on a sound and reliable foundation" (GILBERT, 1971, p. 526).

In the twentieth century the process of completing the story of history has taken two main directions: the establishment of the main features of the history of non-European countries, especially those in the underdeveloped regions of the world, and the accurate portrayal of the life and times—the culture—of those ordinary people whose stories were skipped over, or only lightly touched, by the scholars of the nineteenth century. Much novelty now comes from finding ways of revealing the *mentalité* of pre-industrial peasants in France, the religious beliefs of English workers or U.S. slaves, the evolution of love and the family in Britain, the changing position of the aged in American society, the rise of literacy in Britain or in France, and the movement of women from the household into the labor force.²⁴

The exploitation of "new" evidence is another major source of novelty. By "new" I mean not only recently discovered bodies of evidence but, what in practice has proved to be more important, new attacks on large bodies of long-existing evidence that previously seemed of little relevance or else seemed too massive to be penetrated. Elton's reinterpretation of the politics of the Tudor era, for example, as he has emphasized, turns largely on moving from the information contained in the *Calendar of Letters and Papers of Henry VIII* to the documents described by that calendar. "The provenance of the documents—the way in which they came to be produced and deposited—is one of their most telling aspects, and this is something that, disastrously, cannot be established from that calendar" (Elton, 1971, p. 69). It is in the attack on massive but long-neglected bodies of evidence that traditional and cliometric historians have perhaps their greatest affinity. Despite considerable differences in the nature of the documents that are the focus of each group, and in the methods of extracting information, both groups have displayed a common ingenuity in coping with the sheer bulk of the evidence and

²⁴ See among other studies LADURIE (1974, 1978), THOMPSON (1975), GENOVESE (1974), LEVINE (1977), STONE (1969 and 1977a), FISCHER (1977), FURET and OZOUF (1978), HIGONNET (1978), BROWNLEE (1979).

in making once obscure and irrelevant documents reveal aspects of history never intended for revelation by those who originally produced the documents.

Attitudes toward collaboration. Traditional historians do not usually collaborate, except in the writing of textbooks, where collaboration is common. There have been various multivolume series, such as the *History of American Life*, to which a number of authors each contributed a volume, or such as the various Cambridge histories, to which a larger number of scholars each contributed one or more chapters. But in these cases, each of the volumes or chapters was the personal product of a particular scholar, and not the collective product of the group. The great classics of traditional history have all had a highly personal voice. Just as it is *Shakespeare's "Julius Caesar"*, it is *Gibbon's "Decline and Fall of the Roman Empire,"* and *Prescott's "History of the Conquest of Mexico."* Traditional historians do not consider a highly personal voice to be a failing, a departure from objectivity. Quite the contrary, the quality of the mind and spirit of the author as it emerges from the pages of his history is a central element in the assessment of the history. The traditional historian is expected to draw moral lessons and quite frequently two studies of the same historical question will differ not so much in the statement of the facts as in their moral stance.

While not all cliometric historians are involved in large-scale, collaborative research, such projects are a hallmark of cliometric history. Collaboration is necessary partly because the scope of the data collection requires many hands and partly because no one scholar can be expected to master all of the technical skills required for such projects. In their study of English population and social structure the Cambridge Group, for example, has had to involve not only programmers, statisticians, and mathematicians but also demographers, sociologists, anthropologists, economists, physiologists, nutritionists, geneticists, and epidemiologists.²⁵ The basis for

²⁵ Private communication with Peter Laslett, August 1979. Other examples of large cliometric collaborations in Europe include the programs on Scottish demographic history at Edinburgh and Aberdeen, the project on Swedish family history at Uppsala, the institutes for mathematical studies in history in Moscow and Tallinn, the project on German demographic history at the Free University of Berlin, the maritime history project in England, the anthropological reconstruction of a village in Essex from medieval to modern times, the project on the history of French literacy at the École des Hautes Études en Sciences Sociales in Paris, and the project on the Florentine catasto also sponsored by the École des Hautes Études. Examples of large cliometric collaborations in North America include the colonial history program of the St. Mary's City Commission

collaboration is usually technical, and the moral viewpoints of the collaborators frequently span the entire ideological spectrum. The products of these intellectual enterprises are frequently multiauthored; one recent paper involved no less than nine authors.

Personal voice does not entirely disappear from such works but is quite muted, and many cliometricians treat a marked personal voice as a failing. Since they are concerned with facts and behavioral regularities that can be established objectively, moralizing is considered to be entirely out of place. And when the course of research unavoidably touches on moral issues, cliometricians typically strive to treat these with a detachment and coolness that repels many traditional historians. This air of detachment, even when moral issues are not directly involved, arouses the suspicion of some traditional historians who cannot perceive the moral stance of the investigators and who fear that what is offered as objective evidence, is not, and somehow will turn out to be inimical to their ideological positions. Some traditional historians refer to these collaborations as "factories" and to collaborators as "helots."²⁶ They appear to be unaware of the high degree of intellectual tension that exists among the collaborators in such projects and seem to imagine that like some German seminars of the nineteenth century, they are composed of a domineering professor and his disciples.

Cliometric collaboration is carried on not just within projects, but across projects as well. In traditional history a scholar working on the policies of the administration of Franklin D. Roosevelt would not normally consider the work of a historian of the French revolution relevant to his concerns, except in certain indirect ways.²⁷ But U.S. demographic historians,

in Annapolis, the program on economic factors in the U. S. fertility decline at the University of Pennsylvania, the Philadelphia Social History Project, the European Fertility Project at Princeton University, the Harvard-Brown-Chicago project on medieval Florentine dowries, Programme de recherche en demographie historique in Quebec, the Center for Research on Social Organization at the University of Michigan, the Newberry Family and Community History Center in Chicago, the Mormon Historical Demography Project at the University of Utah, the program on post-Columbian population changes in Mexico and other Latin American regions at the University of California (Berkeley), and the joint Berkeley-Cambridge-Harvard simulations of historical social structures. This listing, it should be added, is far from exhaustive.

²⁶ Friction has also been generated by the large grants that have been awarded for cliometric research. See the exchange between Parker and Haskell in *The New York Review*, December 11, 1975.

²⁷ As when Genovese finds an analogy between the resistance of slave craftsmen and the revolutionary activities of French and British artisans in the eighteenth century. Cf. p. 39 above.

for example, follow research on fertility and mortality in other countries with great intensity, partly because such findings are often directly relevant to their work, partly because it is of immediate importance to compare the findings on demographic behavior in a particular region and time with what has been established about such behavior in other places and times, and partly because the analytical techniques devised by one investigator often are directly applicable to the problems of another. The members of the various cliometric groups keep in close touch with each other, exchange papers, meet personally whenever possible, address one another's seminars on work in progress, and hold frequent conferences to assess the state of the art. Research centers that are technically advanced, such as the Cambridge Group, have attracted scholars throughout Europe and America. The visits are sometimes brief (a few days or a few weeks) but, especially in the case of younger scholars, often run on for an entire year. Although limited in their resources, cliometric centers such as the Cambridge Group, the Philadelphia Social History Project, the Newberry Family and Community History Center, the Economic History Workshop at Chicago, and the Institutes for Social, Family, and Economic History at Uppsala welcome these visitors.

Communication with the history-reading public. Traditional historians place great emphasis on communicating with a public that is wider than themselves. "Historians", said the authors of the *Harvard Guide*, "are not, of course, the sole creators of tradition" for "orators, poets, politicians, clergymen" and many others also contribute to that end. But they believed that "more than any other class of writers or teachers", especially because of "their access to the youthful mind," historians influence "a people's conception of its past." Lawrence Stone goes even further, calling histories "essential elements in creating the high culture of their time" and emphasizing the capacity of "sober apparent truth, as elegantly told by historians" to be "more gripping, more intriguing, and more meaningful" than "artificial romances and novels."²⁸

Cliometrists do not generally address this wider public and are frequently disdainful of any in their number who attempt to do so. Some doubt the wisdom of entering into moral and aesthetic realms and do not feel that historians have either the obligation or a special qualification to be

²⁸ HANDLIN *et al.* (1954, p. 9), STONE (1977b, pp. 3-4). Gilbert points out that, leaving cliometrics aside, there has been a certain weakening of attention to the general audience for history and an increasing tendency for traditional historians to address each other.

the moral guardians of the young. They tend to be disdainful of efforts to reconstruct the "motives and feelings of [particular] long-dead individuals," and some believe that however dramatic and compelling such attempts might be, they are "beyond the reach of empirical inquiry" and are "better left" to the "evocative methods of poets" (CLUBB and BOGUE, 1977, p. 180). Many cliometricians want to concentrate on the production of empirically warranted statements about the past that have direct relevance to present day issues and concerns. Many hope that by studying the past they can discover warranted generalizations about human behavior that have force in the present and will continue to do so in the future. The majority of cliometricians believe that the proper audience for such works are not those who read history for pleasure but those who are capable of assessing and validating the fruits of scientific labors—not a broad public, but a narrow group of highly trained specialists.

Relations between "scientific" and traditional historians

Overall, relationships between the adherents of the two traditions can be characterized as those of cultural warfare. There was, of course, bound to be hostility since one would hardly expect those who have woven the fabric of traditional history to be unmoved by cliometric efforts to tear it to shreds. It was not merely that quantification revealed significant errors in the work of traditional historians; antagonism was also fanned by the extremely aggressive stance of the cliometricians and by exaggerated claims. As David Landes recently put it, cliometric criticism was almost entirely devoid of those courtesies and tokens of respect that soften the edge of criticism and "make even gall moderately palatable". Some cliometricians seemed to believe that the whole of traditional historiography was so laced with error as to be almost wholly useless. The message to them, said one traditional historian, was to "Retool, rethink, reform, or be plowed under" (LANDES, 1978, p. 4; KIRKLAND, 1967, p. 1494).

Traditional historians have generally refused to accede to the cliometricians. Some have simply ignored the challenge. Others have called upon cliometricians to temper their language and claims and have sought to incorporate cliometric findings into the stream of traditional historiography. But the predominant response has been to counterattack in a variety of ways. One thrust has been to acknowledge that traditional history is frequently too impressionistic, too imprecise and could be improved by formalization of arguments and more careful measurement, but to deny

that cliometric efforts along these lines have significantly advanced knowledge, except perhaps in economic history which is treated as a special case. Another line has been to attack the results of cliometric work on the grounds that the bodies of evidence utilized by the cliometricians are worthless, their statistical procedures inappropriate or misapplied, and their analyses logically or ideologically flawed. A third line of attack has been to argue that preoccupation with statistics and behavioral models has given cliometricians such extraordinarily simplistic views of human motivations, relationships, personalities, and ethics that they are incapable of sensibly interpreting their own findings. The charge that cliometricians are too innocent to be allowed to treat questions of psychology and culture is neatly interwoven with suggestions that they slyly manipulate their statistical measures to obtain results that advance their hidden ideological positions.²⁹

The reluctance to acknowledge that cliometric labors have produced important results leads antcliometric warriors to curious positions. Cliometric results are grudgingly accepted but not acknowledged. Some traditional historians, for example, now report that slavery was "obviously" a profitable investment to slaveholders since the value of their estates clearly increased over time—an obvious fact that somehow was less obvious or less conclusive before the long cliometric debate on profitability culminated in a consensus. Nor is it credible that cliometricians who are so naive, simplistic, careless, and sloppy, can so often reach results that cause traditional historians to alter their positions. It is also curious that studies which are instantly declared to be *obviously* wrong, nevertheless call forth paper after paper aimed at disproving them. Rather than being banished, such works become the explicit or implicit points of reference not only for subsequent cliometric research but for traditional research as well. While such tactics may keep some graduate students away from cliometrics, to others it adds the allure of forbidden fruit. Whatever the net result, the fact remains that cliometric influence has increased steadily, especially among young historians who, with or without the approval of their teachers, are struggling to master scientific methods and the art of applying them to history.

Curious positions and grudging concessions are terms that also apply to antitradiional warriors. The anticipated rout of traditional historians

²⁹ COCHRAN (1969), ERICSON (1975), GUTMAN (1976), HANDLIN (1975), HASKELL (1975a and 1975b), HEXTOR (1967), HIGHAM (1970), PARKER (1975), SCHEIBER (1975), STONE (1977b), WOODWARD (1968).

has not materialized and history has not been transformed into a science. Cliometricians have had to acknowledge that there are issues for which traditional methods are better suited than scientific ones. Moreover, successful application of cliometric methods requires a deep and thorough knowledge of historical circumstances. Solid cliometric contributions are generally the product of painstaking searches of archives for primary data, mastery of the secondary literature, and immersion in the public and private documents. Such work is a precondition to the successful application of powerful general methods to specific historical situations. No amount of mathematical wizardry or computer magic can shortcut this process. Efforts to do so have led to embarrassing failures.

Recognition that behavioral models must be place- and time-specific to be useful in historical research differentiates cliometricians from the scientific historians of the late nineteenth century who hoped to discover generalizations that were truly timeless—that would be equally valid for Babylon, for Victorian Britain, and for America in the twenty-first century. Cliometricians have discovered that few generalizations cover such vast stretches of human experience and those that do are so vague as to be of little operational value. Experience has revealed that regularities which can be estimated by cliometric methods, such as demand curves for particular commodities, equations that relate fertility to social and economic variables, or equations that describe political behavior, are not the same for all times and places but differ in varying degrees from time to time or place to place. Far from diminishing historical specificity, cliometric techniques have often shown that processes that seemed continuous over time were actually quite discontinuous, and behavior that seemed similar in two places was actually quite dissimilar. Far from abstracting from details, cliometrics has led to ever deeper probing into the details of the slave system, family structure, social mobility, ethnic influences on popular voting behavior, and a whole gamut of other issues that have come under cliometric scrutiny.

Even in American economic history, where it is generally agreed that scientific methods have had their greatest impact, cliometric findings do not yet add up to a continuous account of economic development over the last 370 years. The new economic historians have made important, sometimes even far-reaching contributions, but these have been on quite specific points within the traditional account. They may have altered, but they have not replaced, the basic narratives of the growth of agriculture, the rise of manufacturing, the evolution of banking, the spread of trade, and much else that has been traced and documented by traditional methods. Some

cliometrists argue that their contributions, though confined to a limited set of points, have placed major parts of the traditional narrative in a new light and have extended the narrative in novel directions. The point, however, is that whether cliometrics opens up new avenues of knowledge, overturns particular elements in the traditional narrative, or merely refines some elements, its contribution is to the elaboration of the narrative. Cliometrics has not made narrative history obsolete.

The genuine differences between "scientific" and traditional historians over subject matter, methods, and style should not obscure their more numerous and more fundamental affinities and complementarities. There is much wisdom in Elton's observation that the basic methodological approaches to history (not the exact details of techniques) "were worked out quite early." Cliometrists who doubt it should turn to Book I, part 10, of *Peloponnesian Wars* to see how close Thucydides was to their style (or they to his) when he estimated the size of the assault force on Troy, a point that is critical in establishing his contention that the Peloponnesian wars overshadowed all previous ones in Greek history. Traditional historians and others who doubt the possibility of blending poetics and counting should turn to Lincoln's Cooper Institute address (February 27, 1860) for a magnificent mid-nineteenth century example of such a blend.³⁰

The grudging concessions on both sides are tacit admissions that neither mode of research by itself is adequate to deal with all of the questions that concern historians. To explain the outbreak of the Civil War, for example, one must deal not only with systematic forces in the economic, social, ideological, and political spheres that may have made such a crisis likely, but also with the role of particular personalities, unique events, decisions that could well have gone differently, blunders, and a host of other contingent factors that loomed large in the actual course of events. Thus, while "scientific" and traditional history are different and, in some respects, competing modes of research, they are neither mutually exclusive nor intrinsically antagonistic. Quite the contrary, precisely because each mode has a comparative advantage in certain domains of research, they supplement and enrich each other. It seems reasonable to believe that as the tendency toward the interpenetration of the two modes continues, the intensity of the cultural conflict between them will diminish. Those who have studied cultural wars know, however, that irrational factors often have as much to do with the course of such conflicts as rational ones. Is it self-delusion

³⁰ ELTON (1967, p. 6). I am indebted to Professor Ephim G. Fogel of Cornell University for many helpful suggestions regarding the presentation of the material in this paper.

to assume that historians are more likely to benefit from the lessons of history than other folk? I cannot free myself of the belief that in this case rationality will prevail.

References

- ADAMS, George B., 1909, *History and the philosophy of history*, American Historical Review, vol. 14, pp. 221–236
- ADAMS, Henry, 1894, *The tendency of history*, Annual Report of the American Historical Association, pp. 17–23
- ADAMS, Herbert B., 1895, *Is history past politics?* Johns Hopkins Univ. Studies, vol. 13, pp. 67–81
- ALCORN, R. S., and Peter R. KNIGHTS, 1975, *Most uncommon Bostonians: A critique of Stephan Thernstrom's 'The other Bostonians'*, Historical Methods, vol. 8, pp. 98–114
- ANDREANO, Ralph, ed., 1970, *The new economic history: Recent papers in methodology* (New York)
- AYDELOTTE, W. O., 1966, *Quantification in history*, American Historical Review, vol. 71, pp. 803–825
- AYDELOTTE, W. O., 1971, *Quantification in history* (Reading, Mass.)
- AYDELOTTE, W. O., 1973, *Lee Benson's scientific history: For and against*, Journal of Interdisciplinary History, vol. 4, pp. 263–372
- AYDELOTTE, W. O., et al., 1972, *The dimensions of quantitative research in history* (Princeton)
- AYDELOTTE, W. O., ed., 1977, *The history of parliamentary behavior* (Princeton)
- AYMARD, M., 1971, *The Annales and French historiography (1929–72)*, The Journal of European Economic History, vol. 1, pp. 491–511
- BAILYN, B., 1977, *Review article, 'French Historical Method: The Annales Paradigm'*, by Traian Stoianovich, Journal of Economic History, vol. 37, pp. 1028–1034
- BALLARD, M., ed., 1970, *New movements in the study and teaching of history* (Bloomington, Indiana)
- BARNES, H. E., 1962, *A history of historical writing*, 2nd rev. ed. (New York)
- BENSON, L., 1960, *Turner and Beard: American historical writing reconsidered* (Glencoe, Ill.)
- BENSON L., 1972, *Toward the scientific study of history* (Philadelphia)
- BERMAN, H. J., 1968, *Legal reasoning*, in: *International Encyclopedia of the Social Sciences*, New York, vol. 9, 197–204
- BOGUE, A. G., 1968, *United States: The 'new' political history*, Journal of Contemporary History, vol. 3, pp. 5–27
- BOGUE, A. G., 1976, *Comment on paper by Easterlin*, Journal of Economic History, vol. 36, pp. 76–81
- BOGUE, A. G., 1978, *Recent developments in political history: The case of the United States*, in: *The Frontiers of Human Knowledge* (Uppsala), pp. 79–109
- BRAYBROOK, D., 1977, *Review essay, 'Casual explanation and model building in history... by Peter D. McClelland*, History and Theory, vol. 16, pp. 337–354
- BRINTON, C., 1930, *The Jacobins: An essay in the new history* (New York)

- BRINTON, C., 1930, *The new history: Twenty-five years after*, Journal of Social Philosophy, vol. 1, pp. 134–147
- BROWNLEE, W. E., 1979, *Household values, women's work, and economic growth*, Journal of Economic History, vol. 39, pp. 199–209
- BURGESS, J. W., 1896, *Political science and history*, Annual Report of the American Historical Association, pp. 203–219
- CLAPHAM, Sir John, 1931, *Economic history as a discipline*, in: *Encyclopedia of the Social Sciences*, New York, vol. V, pp. 327–330
- CLUBB, J. M., 1975, *The 'new' history as applied social science: A review essay*, Computers and the Humanities, vol. 9, pp. 247–251
- CLUBB, J. M., and H. W. ALLEN, 1967, *Computers and historical studies*, Journal of American History, vol. 54, pp. 599–607
- CLUBB, J. M., and H. W. ALLEN, 1977, *Collective biography and the progressive movement: The 'status revolution' revisited*, Social Science History, vol. 1, pp. 518–534
- CLUBB, J. M., and A. G. BOGUE, 1977, *History, quantification, and the social sciences*, American Behavioral Scientist, vol. 21, pp. 167–186
- COCHRAN, T. C., 1969, *Economic history, old and new*, American Historical Review, vol. 74, pp. 1561–1572
- COLEMAN, D. C., 1977, *The model game*, Economic History Review, vol. 30, pp. 346–351
- CONRAD, A. H., and J. R. MEYER, 1964, *The economics of slavery and other studies in econometric history* (Chicago)
- CRATON, M., 1978, *Searching for the invisible man: slaves and plantation life in Jamaica* (Cambridge, Mass.)
- CRAWFORD, S. C., 1980, *The autobiography of an oppressed class*, Ph. D. thesis, Univ. of Chicago
- DAVID, P. A., et al., 1976, *Reckoning with slavery* (New York)
- DAVID, P. A., and P. TEMIN, 1979, *Explaining the relative efficiency of slave agriculture in the United States: Comment*, American Economic Review, vol. 69, pp. 213–218
- DAVIS, L. E., 1968, 'And it will never be literature': *The new economic history: A critique*, Explorations in Economic History (Fall), vol. 6, pp. 75–92
- DAVIS, L. E., 1971, *Specification, quantification, and analysis in economic history*, in: TAYLOR and ELLSWORTH, pp. 106–120
- DELZELL, Ch. F., 1977, *The future of history* (Nashville)
- DOLLAR, Ch. M., and R. J. JENSEN, 1971, *Historian's guide to statistics: Quantitative analysis and historical research* (New York)
- DRAKE, M., ed., 1969, *Population in industrialization* (London)
- DROYSSEN, J. G., 1893, *Outline of the principles of history* (Boston)
- EASTERLIN, R. A., 1961, *Regional income trends, 1840–1950*, in: American economic history, ed. S. Harris (New York)
- EASTERLIN, R. A., 1976, *Population change and farm settlement in the Northern United States*, Journal of Economic History, vol. 36, pp. 45–75
- EASTERLIN, R. A., 1977, *Population issues in American economic history: A survey and critique*, in: Recent developments in the study of business and economic history, research in economic history, ed. R. E. Gallman, Supplement 1, pp. 131–158
- ELSTER, J., 1978, *Logic and society: Contradictions and possible worlds* (Chichester)
- ELTON, G. R., 1967, *The practice of history* (Sydney)

- ENGERMAN, S. L., 1975, *Up or out: Social and geographic mobility in the United States*, *Journal of Interdisciplinary History*, vol. 5, pp. 469–489
- ENGERMAN, S. L., 1977, *Recent developments in American economic history*, *Social Science History*, vol. 2, No. 1, pp. 72–89
- ENGERMAN, S. L., 1980, *Counterfactuals and the new economic history*, *Inquiry*, vol. 23, pp. 157–172
- ENGERMAN, S. L., and E. D. GENOVESE, eds., 1975, *Race and slavery in the western hemisphere: quantitative studies* (Princeton)
- ERICSON, Ch., 1975, *Quantitative history*, *American Historical Review*, vol. 80, pp. 351–365
- FIELD, D., 1969, *I. D. Kovalchenko: Russkoe krest'ianstvo v pervoi polovine XIX veka*, *Kritika*, vol. 5, pp. 31–45
- FINLEY, M. I., 1977, 'Progress' in historiography, *Daedalus*, Summer, vol. 106, pp. 125–142
- FISCHER, D. H., 1977, *Growing old in America* (New York)
- FISHLOW, A., 1964, *Antebellum interregional trade reconsidered*, *American Economic Review*, vol. 59, pp. 352–364
- FLINN, M. W., 1970, *British population growth, 1700–1850* (London)
- FLINN, M. W., et al., 1977, *Scottish population history, from the 17th Century to the 1930s* (Cambridge)
- FLOUD, R., 1973, *An introduction to quantitative methods for historians* (London)
- FOGEL, R. W., 1965, *The reunification of economic history with economic theory*, *American Economic Review*, vol. 55, pp. 92–98
- FOGEL, R. W., 1967, *The specification problem in economic history*, *Journal of Economic History*, vol. 27, pp. 283–308
- FOGEL, R. W., 1970, *History and retrospective econometrics*, *History and Theory*, vol. 9, No. 3, pp. 245–264
- FOGEL, R. W., 1975, *The limits of quantitative methods in history*, *American Historical Review*, vol. 80, pp. 329–350
- FOGEL, R. W., 1979, *Notes on the social saving controversy*, *Journal of Economic History*, vol. 39, pp. 1–54
- FOGEL, R. W., and S. L. ENGERMAN, 1974, *Time on the cross*, 2 vols. (Boston)
- FOGEL, R. W., and S. L. ENGERMAN, 1977, *Explaining the relative efficiency of slave agriculture in the antebellum South*, *American Economic Review*, vol. 67, pp. 275–296
- FOGEL, R. W., and S. L. ENGERMAN, 1980, *Explaining the relative efficiency of slave agriculture in the antebellum South: Reply*, *American Economic Review*, vol. 70, pp. 672–690
- FORSTER, R., 1978, *The achievements of the Annales School*, *Journal of Economic History*, vol. 38, pp. 58–76
- FURET, F., 1971, *Quantitative history*, *Daedalus*, Winter, vol. 100, pp. 151–167
- FURET, F., and E. LE ROY LADURIE, 1970, *L'historian et l'ordinateur: Compte-rendu provisoire d'enquête*, Rapport collectif présenté par le Centre de Recherches Historiques de l'Ecole Pratique des Hautes Etudes (Moscow)
- FURET, F., and J. OZOUF, eds., 1978, *L'alphabétisation des Français de Calvin à Jules Ferry*, 2 vols. (Paris)
- GALLMAN, R. E., 1971, *The statistical approach: Fundamental concepts as applied to history*, in: TAYLOR and ELLSWORTH, eds., pp. 63–86
- GALLMAN, R. E., 1979, *Slavery and Southern economic growth*, *Southern Economic Journal* vol. 45, pp. 1007–1022

- GARDINER, P., 1968, *History: The philosophy of history*, International Encyclopaedia of the Social Sciences, New York, vol. 6, pp. 428-434
- GAY, P., 1974, *Style in History* (New York)
- GENOVESE, E. D., 1974, *Roll Jordan Roll: The world the slaves made* (New York)
- GENOVESE, E. D., and E. FOX-GENOVESE, 1979, *The slave economies in political perspective*, Journal of American History, vol. 66, pp. 7-23
- GILBERT, F., 1971, *Post scriptum*, Daedalus, Spring, vol. 100, pp. 520-530
- GLASS, D. V., and D. E. C. EVERSLY, eds., 1965, *Population in History: Essays in historical demography* (Chicago)
- GOULD, J. D., 1969, *Hypothetical history*, Economic History Review, vol. 22, pp. 195-207
- GRAY, L. C., 1933, *History of agriculture in the Southern United States to 1860*, 2 vols. (Washington, D. C.)
- GUTMAN, H. G., 1975, *Slavery and the numbers game: A critique of time on the cross* (Urbana, Ill.)
- GUTMAN, H. G., 1976, *The black family in slavery and freedom, 1750-1925* (New York)
- HABAKKUK, J., 1971, *Economic history and economic theory*, Daedalus, Spring, vol. 100, pp. 305-322
- HANDLIN, O., 1941, *Boston's immigrants, 1790-1880* (Cambridge, Mass.)
- HANDLIN, O., 1975, *The capacity of quantitative history*, Perspectives in American History, vol. 9, pp. 7-26
- HANDLIN, O., 1979, *Truth in history* (Cambridge, Ma.)
- HANDLIN, O., et al., 1954, *Harvard guide to American history* (Cambridge, Ma.)
- HARAVEN, T., 1977, *Family time and historical time*, Daedalus, Spring, vol. 106, pp. 57-70
- HARAVEN, T., and M. VINOVSKIS, eds., 1978, *Family and population in nineteenth-century America* (Princeton)
- HASKELL, T. L., 1975a, *The true and tragical history of 'Time on the cross'*, The New York Review, October 2, pp. 33-38
- HASKELL, T. L., 1975b, *Funds for Clio: Thomas L. Haskell replies*, The New York Review, December 11, p. 61
- HARTWELL, R. M., 1971, *Is the new economic history an export product?*, A comment on J. R. T. Hughes, in: McCLOSKEY, ed., pp. 413-422
- HAWKE, G. R., 1970, *Railways and economic growth in England and Wales 1840-1870* (Oxford)
- HAYS, S. P., 1968, *New possibilities for American political history: The social analysis of political life*, in: LIPSET and HOFSTADTER, eds., pp. 181-227
- HECKSHER, E. F., 1929, *A plea for theory in economic history*, Economic History 1, Supplement to the Economic Journal, January, pp. 525-534
- HERBST, J., 1965, *The German historical school in American scholarship* (Ithaca)
- HERSHBERG, T., et al., 1976, *A special issue*, *The Philadelphia social history project*, Historical Methods Newsletter, vol. 9, pp. 41-181
- HEXTER, J. H., 1967, *Some American observations*, Journal of Contemporary History, vol. 2, pp. 5-23
- HEXTER, J. H., 1972, *Fernand Braudel and the Monde Braudellien...*, Journal of Modern History, vol. 44, pp. 480-539
- HIGHAM J., 1970, *Writing American history: Essays on modern scholarship* (Bloomington, Indiana)
- HIGHAM, J., et al., 1965, *History* (Englewood Cliffs, N. J.)

- HIGMAN, B. W., 1976, *Slave population and economy in Jamaica* (Cambridge)
- HIGMAN, B. W., 1978, *African and creole slave family patterns in Trinidad*, Journal of Family History, vol. 3, pp. 163–180
- HIGONNET, P., 1978, *Reading, writing, and revolution*, Times Literary Supplement, October 13, p. 1153
- HILL, C., 1978, *Sex, marriage, and the family in England*, Economic History Review, August, vol. 31, pp. 450–463
- HOFSTADTER, R., 1955, *Social darwinism in American thought*, rev. ed. (Boston)
- HOFSTADTER, R., 1956, *History and the social sciences*, in: STERN, ed.
- HOFSTADTER, R., 1968a, *The progressive historians: Turner, Beard, Parrington* (New York)
- HOFSTADTER, R., 1968b, *History and sociology in the United States*, in: LIPSET and HOFSTADTER, ed.
- HOLT, W. S., 1940, *The idea of scientific history in America*, Journal of the History of Ideas, vol. 1, pp. 352–362
- HUGHES, H. S., 1958, *Consciousness and society: The reorientation of European social thought 1890–1930* (New York)
- HUGHES, J. R. T., 1966, *Fact and theory in economic history*, Explorations in Entrepreneurial History, 2nd Ser., vol. 3, pp. 75–100
- HUGHES, J. R. T., 1971, *Is the new economic history an export product?*, in: McCLOSKEY, ed., pp. 401–412
- HULL, Ch. H., 1914, *The service of statistics to history*, Quarterly Publications of the American Statistical Association, March, vol. 14, pp. 30–39
- IGGERS, G. G., 1962, *The image of Ranke in American and German historical thought*, History and Theory, vol. 2, No. 1, pp. 17–40
- IMHOF, A. E., 1978a, *The analysis of eighteenth-century causes of death: Some methodological considerations*, Historical Methods, vol. 11, pp. 3–35
- IMHOF, A. E., 1978b, *The computer in social history: Historical demography in Germany*, Computers and the Humanities, vol. 12, pp. 227–236
- JENSEN, R., 1974, *Quantitative American studies: The state of the art*, American Quarterly, vol. 26, pp. 225–240
- JOYNT, C. B., and N. RESCHER, 1961, *The problem of uniqueness in history*, History and Theory, vol. 1, pp. 150–162
- KAHK, J., and I. D. KOVALCHENKO, 1974, *Methodological problems of the application of mathematical methods in historical research*, Historical Methods, June, vol. 7, pp. 217–224
- KEMBLE, F. A., 1961, *Journal of a residence on a Georgian plantation 1838–1839*, edited with an introduction by John A. Scott (New York)
- KIRKLAND, E. C., 1967, *Review of "Railroads and American Economic Growth" by Robert W. Fogel*, American Historical Review, vol. 72, pp. 1493–1495
- KOVALCHENKO, I. D., 1967, *Russkoe krepostnoe krest'ianstvo v pervoi polovine XIX v.* (The enserfed peasantry of Russia in the first half of the nineteenth century, Russian) (Izdatel'stvo Moskovskogo universiteta, Moscow)
- KOVALCHENKO, I. D., et al., eds., 1977, *Mathematical methods in historical-economic and historical cultural studies* (published in Russian by NAUKA, Moscow)
- KOUSSEY, J. M., 1976, *The new political history: a methodological critique*, Reviews in American History, March, vol. 4, pp. 1–14

- KOSSER, J. M., 1977, *The agenda for 'Social science history'*, Social Science History, vol. 1, pp. 383-391
- KOSSER, J. M., 1979, *Quantitative social scientific history*, in: M. Kammen, ed., *The Past Before US: Contemporary historical writing in the United States* (Ithaca), pp. 433-456
- KREIGER, L., 1977, *Ranke: The meaning of history* (Chicago)
- KUHN, T. S., 1970, *The structure of scientific revolutions*, 2nd ed. (Chicago)
- LE ROY LADURIE, E., 1974, *The peasants of Languedoc* (Urbana)
- LE ROY LADURIE, E., 1978, *Montaillou: The promised land of error* (New York)
- LE ROY LADURIE, E., 1979, *The territory of the historian* (Hassocks, Sussex, England)
- LAMPARD, E. E., 1975, *Two cheers for quantitative history: An agnostic forward*, in: SCHNORE
- LANDES, D. S., 1978, *On avoiding Babel*, Journal of Economic History, vol. 38, pp. 3-12
- LANDES, D. S., and Ch. TILLY, 1971, *History as social science* (Englewood Cliffs, N. J.)
- LASLETT, P., ed., 1972, *Household and family in past time* (Cambridge)
- LASLETT, P., 1977, *Family life and illicit love in earlier generations* (Cambridge)
- LEE, R. D., 1977, *Population patterns in the past* (New York)
- LEVINE, L. W., 1977, *Black culture and black consciousness* (New York)
- LE GOFF, J., 1971, *Is politics still the backbone of history?*, Daedalus, Winter, vol. 100, pp. 1-19
- LÉVY-LEBOYER, M., 1969, *La 'New economic history'*, Annales, vol. 24, pp. 1035-1069
- LIPSET, S. M., and R. HOFSTADTER, eds., 1968, *Sociology and history: methods* (New York)
- LOCKRIDGE, L. A., 1977, *Historical demography*, in: DELZELL, 1977 (pp. 53-64)
- LORWIN, V. R., and J. M. PRICE, eds., 1972, *The dimensions of the past: materials, problems, and opportunities for quantitative work in history* (New Haven)
- MACFARLANE, A., 1979, *Review of 'The family, sex, and marriage 1500-1800'*, by Lawrence Stone, History and Theory, vol. 18, pp., 103-126
- MCCORMICK, R. L., 1974, *Ethno-cultural interpretations of nineteenth century American voting behavior*, Political Science Quarterly, vol. 89, pp. 351-377
- MCCLELLAND, P., 1975, *Causal explanation and model building in history, economics, and the new economic history* (Ithaca, N. Y.)
- MCCLELLAND, P., 1978, *Cliometrics versus institutional history*, Research in Economic History, vol. 3, pp. 369-378
- MCCLOSKEY, D. N., ed., 1971, *Essays on a mature economy* (Princeton)
- MCCLOSKEY, D. N., 1978, *The achievements of the Cliometric School*, Journal of Economic History, vol. 38, pp. 13-28
- MCKEOWN, T., 1976, *The modern rise of population* (New York)
- MARWICK, A., 1976, *The nature of history* (London)
- MATHIAS, P., 1970, *Economic history—direct and oblique*, in: BALLARD, 1970
- METZER, J., 1977, *Some economic aspects of railroad development in Tsarist Russia* (New York)
- MILOV, L. V., and K. V. KHOVOSTOVA, 1973, *Quantitative methods applied by Soviet historians to agrarian history*, International Conference on the Application of Mathematical Methods in Historical Research, Sweden, June, mimeo
- MURPHY, G. G. S., 1969, *On counterfactual propositions*, History and Theory, Beiheft 9, pp. 14-38
- O'BRIEN, P., 1977, *The new economic history of railroads* (London)

- NORTH, D. C., 1963, *Quantitative research in American economic history*, American Economic Review, vol. 53, pp. 128-130
- NORTH, D. C., 1968, *History: Economic history*, International Encyclopedia of the Social Sciences, New York, vol. 6, pp. 468-474
- NORTH, D. C., 1977, *The new economic history after twenty years*, American Behavioral Scientist, vol. 21, pp. 187-200
- NORTH, D. C., 1978, *Structure and performance: The task of economic history*, Journal of Economic Literature, vol. 16, pp. 963-978
- PARKER, W. N., 1972, *Economic history: Two papers on the development and state of the art*, mimeo, University of Chicago Workshop in Economic History
- PARKER, W. N., 1975, *Funds for Clio*, The New York Review, December 11, p. 61
- PHILLIPS, U. B., 1918, *American Negro slavery* (New York)
- POPE, C. L., 1975, *The impact of the ante-bellum tariff on income distribution* (New York)
- POTTER, J., 1965, *The growth of population in America, 1700-1860*, in: GLASS and EVERSLY, pp. 631-688
- PRICE, J. M., 1969, *Recent quantitative work in history: A survey of the main trends*, History and Theory, Beiheft 9: Studies in quantitative history and the logic of the social sciences, pp. 1-13
- RABB, T. K., and R. I. ROTBERG, eds., 1973, *The family in history: Interdisciplinary perspectives* (New York)
- RAZZELL, P. E., 1974, *An interpretation of the modern rise of population in Europe—A critique*, Population Studies, vol. 28, pp. 5-17
- REDLICH, F., 1965, *New and traditional approaches to economic history and their interdependence*, Journal of Economic History, vol. 25, pp. 480-495
- ROBINSON, J. H., 1912, *The new history: Essays illustrating the modern historical outlook* (New York)
- ROOSEVELT, THEODORE, 1913, *History as literature*, American Historical Review, vol. 18, pp. 473-489
- ROSTOW, W. W., 1948, *British economy of the nineteenth century: Essays* (Oxford)
- ROTHSTEIN, M., et al., 1970, *Quantification and American history: An Assessment*, in: *The state of American history*, ed. Herbert J. Bass (Chicago)
- ROWNEY, D. K., and J. Q. GRAHAM, Jr., eds., 1969, *Quantitative history* (Homewood, Ill.)
- SAVETH, E., 1960, *Scientific history in America: Eclipse of an idea*, in: *Essays in American Historiography*, eds. Donald Sheehan and Harold C. Syrett (New York)
- SAVETH, E., ed., 1964, *American history and the social sciences* (New York)
- SCHEIBER, H. N., 1975, *Black is computable*, The American Scholar, vol. 44, pp. 656-673
- SCHLESINGER, A. M., and D. R. FOX, eds., 1927-1948, *History of American life*, 12 vols., N.Y.
- SCHLESINGER, A., Jr., 1962, *The humanist looks at empirical social research*, American Sociological Review, vol. 27, pp. 768-771
- SCHNORE, L. F., ed., 1975, *The new urban history* (Princeton)
- SCHOFIELD, R. S., 1973, *Dimensions of illiteracy, 1750-1850*, Explorations in Economic History, vol. 10, pp. 437-454
- SCHOFIELD, R. S., and E. A. WRIGLEY, 1980, *English population change* (London)
- SILBY, J. H., 1972, *Clio and computers: Moving into phase II, 1970-1972*, Computers and the Humanities, vol. 7, pp. 67-79
- SILBY, J. H., et al., eds., 1978, *The history of American electoral behavior* (Princeton)

- SIMON, H. A., and N. RESCHER, 1966, *Cause and counterfactual*, Philosophy of Science, vol. 33, pp. 323-340
- SMITH, D. S., 1977, *A homeostatic demographic regime: Patterns in West European family reconstitution studies*, in: LEE, 1977, pp. 19-51
- SMITH, D. S., 1979, *The estimates of early American historical demographers: Two steps forward, one step back, what steps in the future?*, Historical Methods, vol. 12, pp. 24-38
- SPRAGUE, D. N., 1978, *A quantitative assessment of the quantification revolution*, Canadian Journal of History, vol. 13, pp. 177-192
- STAMPP, K. M., 1956, *The peculiar institution: Slavery in the ante-bellum South* (New York)
- STECKEL, R., 1977, *The economics of U. S. slave and southern white fertility*, Ph.D. thesis, University of Chicago
- STERN, F., ed., 1956, *The varieties of history* (New York)
- STOIANOVICH, T., 1976, *French historical method: The Annales paradigm* (Ithaca)
- STONE, L., 1948, *The anatomy of Elizabethan aristocracy*, Economic History Review, vol. 18, pp. 1-53
- STONE, L., 1952, *The Elizabethan aristocracy — a restatement*, Economic History Review, vol. 4 (2nd. Ser.), pp. 302-321
- STONE, L., 1969, *Literacy and education in England, 1640-1900*, Past and Present, vol. 42, pp. 69-139
- STONE, L., 1977a, *The family, sex and marriage 1500-1800* (New York)
- STONE, L., 1977b, *History and the social sciences in the twentieth century*, in: DELZELL, 1977
- STROUT, C., 1966, *The pragmatic revolt in American history: Carl Becker and Charles Beard* (Ithaca)
- SUTCH, R., 1975, *The breeding of slaves for sale and the westward expansion of slavery, 1850-1860*, in: ENGERMAN and GENOVESI, 1975, pp. 173-210
- SWIERENGA, R. P., ed., 1970, *Quantification in American history* (New York)
- SWIERENGA, R. P., 1974, *Computers and American history: The impact of the 'new' generation*, Journal of American History, vol. 60, pp. 1045-1070
- TAYLOR, A. J., ed., 1975, *The standard of living in Britain in the industrial revolution* (London)
- TAYLOR, G. R., ed., 1956, *The Turner thesis concerning the role of the frontier in American history*, rev. ed. (Boston)
- TAYLOR, G. R., and L. F. ELLSWORTH, eds., 1971, *Approaches to the study of American economic history* (Charlottesville, Va.)
- TEMIN, P., 1966. *In pursuit of the exact*, Times Literary Supplement, July 28, pp. 652-653
- THERNSTROM, S., 1973, *The other Bostonians: Poverty and progress in American metropolis, 1880-1970* (Cambridge, Mass.)
- THERNSTROM, S., 1975, *Rejoinder to Alcorn and Knights*, Historical Methods, vol. 8, pp. 115-120
- THOMAS, K. V., 1977, *The changing family*, Times Literary Supplement, October 21, pp. 1226-1227
- THOMPSON, E. P., 1975, *The making of the English working class* (Harmondsworth)
- TILLY, Ch., ed., 1975, *The formation of national states in Western Europe* (Princeton)
- TILLY, Ch., ed., 1978, *Historical studies of changing fertility* (Princeton)
- TILLY, Ch., et al., 1975, *The rebellious century, 1830-1930* (Cambridge, Mass.)

- TREVOR-ROPER, H. R., 1951, *The Elizabethan aristocracy: An anatomy anatomized*, Economic History Review, vol. 3 (2nd. Ser.), pp. 279-298
- TRUSSELL, J., and R. STECKEL, 1978, *The age of slaves at Menarche and their first birth*, Journal of Interdisciplinary History, vol. 8, pp. 477-505
- VANN, R. T., 1969, *History and demography*, History and Theory, Beiheft 9: Studies in Quantitative History and the Logic of the Social Sciences, pp. 64-78
- VINOVSKIS, M. A., 1977, *From household size to the life course: Some observations on recent trends in family history*, American Behavioral Scientist, vol. 21, pp. 263-288.
- VINOVSKIS, M. A., 1978, *Recent trends in American historical demography: Some methodological and conceptual considerations*, Annual Review of Sociology, vol. 4, pp. 603-627
- WACHTER, K. W., et al., 1978, *Statistical studies of historical social structure* (New York)
- WILLIAMSON, J. G., 1974, *Late nineteenth-century American development: A general equilibrium history* (Cambridge)
- WOODMAN, H. D., 1963, *The profitability of slavery: A historical perennial*, Journal of Southern History, vol. 29, pp. 303-325
- WOODMAN, H. D., 1972, *Economic history and economic theory: The new economic history in America*, Journal of Interdisciplinary History, vol. 3, pp. 322-350
- WOODWARD, C. V., 1968, *History and the third culture*, Journal of Contemporary History, vol. 3, pp. 23-25
- WRIGLEY, E. A., 1970, *Population, family and household*, in: BALLARD, 1970
- WRIGLEY, E. A., ed., 1972, *Nineteenth century society: Essays in the use of quantitative methods for the study of social data* (Cambridge)
- WRIGLEY, E. A., and R. S. SCHOFIELD, 1980, *The population history of England, 1541-1871: a reconstruction* (London)
- WRIGHT, G., 1978, *The political economy of the cotton South* (New York)
- WRIGHT, G., 1979, *The efficiency of slavery: Another interpretation*, American Economic Review, vol. 69, pp. 219-226
- ZEISEL, H., 1968, *Statistics as legal evidence*, International Encyclopedia of the Social Sciences, New York, vol. 15, pp. 246-250

THE RÔLE OF MATHEMATICS IN ECONOMICS

WERNER HILDENBRAND

University of Bonn, F.R.G.

Some historical remarks

The use of mathematics in theoretical economics is not at all a recent development, though admittedly classical political economy of the eighteenth and early nineteenth century—a branch of moral philosophy—has been developed and formulated without the use of mathematics. For example, the great masters of classical political economy, Adam Smith and David Ricardo, used only numerical examples to illustrate their theory. They combined observation of facts in an essentially literary manner with deductive analysis of cause and effect relationship to explain the workings of the economic system in which they lived. Even in the writings of John Stuart Mill—the last of the great classical political economists—as well as in the works of Karl Marx, mathematical formulas or diagrams have been used as a shorthand language or an expository device only. The situation drastically changed, however, with the contributions of Cournot, Jevons, Menger, Walras, Gossen, Edgeworth, Marshall, Pareto, etc., that is to say, with the appearance of what we today call *Neoclassical Economics*.

Both schools have been concerned with the production, distribution, exchange and consumption of wealth (wealth mainly as material goods), but the topic on which interest centered, shifted. The classical economists were interested in the changes in the production and distribution of wealth *over time*. In particular, they emphasized the relative rates of population growth and material resources—and examined the consequences of these upon economic progress and the welfare of individuals and society. Neoclassical economists concentrated less on the dynamic feature; they asked: given an economy with a certain population having given tastes, resources and techniques—how can these resources be allocated through a market-

system so as to maximize the satisfaction of consumers? In present day terminology, the transition from classical political economy to neoclassical economics was a shift from macro economic to micro economic analysis. In micro economics the behavior of individual agents is the underlying principle upon which the theory is built. This new orientation—the individual decision problem which was conceived as a maximization problem—could quite naturally be treated mathematically by means of calculus.

Many ardent opponents of the use of mathematics in the social sciences—in the previous century and even today—base their critique on the false claim that mathematics can only be used in areas where all variables can be measured numerically and where the functional relationships can be specified analytically.

A typical statement of this type is:¹ *We can tell that one pleasure is greater than another, but that does not help us. To apply the mathematical methods, pleasure must be in some way, capable of numerical expression. We must be to say, for example, that the pleasure of eating a beef-steak is to the pleasure of drinking a glass of beer as five to four.*

This misconception of mathematics is all the more astonishing, since COURNOT, the recognized father of mathematical economics, wrote in the preface of his *Recherches sur les principes mathématiques de la théorie des richesses* as early as 1838 (COURNOT, pp. 2–3):

“I have said that most authors who have devoted themselves to political economy seem to have had a wrong idea of the nature of the applications of mathematical analysis to the theory of wealth. They imagined that the use of symbols and formulas could only lead to numerical calculations, and as it was clearly perceived that the subject was not suited to such a numerical determination of values by means of theory alone, the conclusion was drawn that the mathematical apparatus, if not liable to lead to erroneous results, was at least idle and pedantic.

But those skilled in mathematical analysis know that its object is not simply to calculate numbers, but that it is also employed to find the relations between magnitudes which cannot be expressed in numbers and between functions whose law is not capable of algebraic expression.”

Many other mathematical economists made similar statements. For example, EDGEWORTH wrote in his *Mathematical Psychics* (1881) an appendix with the title *On Unnumerical Mathematics*. He tried hard to convince his contemporaries not only with mathematical arguments but even by

¹ Quoted from a review of JEVONS, 1871.

resorting to fine prose: *We cannot count the golden sands of life; we cannot number the 'innumerable smile' of seas of love; but seem capable of observing that there is here a greater, there a less multitude of pleasure units, mass of happiness; and that is enough* (EDGEWORTH, 1881, p. 8f).

The narrow view of the nature of mathematics has, by this time, definitely changed. However, many economists are even today of the opinion that for mathematics to be useful, one must be able to calculate the solution explicitly; a view which quite often obliges them to make unnecessary and unjustifiable *ad hoc* assumptions as to the relationships between economic variables.

Most mathematical economists of the last century tried to defend and justify the use of mathematics. Walras—probably the greatest mathematical economist of the previous century—who commented a great deal upon the subject of this lecture in his *Éléments D'Économie Politique Pure*, wrote for example, (WALRAS, 1900, p. 144)

D'où il ressort en fin de compte que la forme mathématique est pour l'économie politique pure non seulement une forme possible, mais la forme nécessaire et indispensable. Je pense, au surplus, que c'est là un point à l'égard duquel aucun des lecteurs qui m'auront suivi jusqu'ici ne saurait conserver le moindre doute.

In retrospect, it seems to me not exaggerated to claim that over the last hundred years the most important contributions to economic theory have been made by mathematically minded persons—I did not say mathematicians! These economists used mathematics more or less extensively, yet they all considered mathematics as necessary and indispensable. There are, however, two famous exceptions which we ought to mention in this context—the two Englishmen—Marshall and Keynes. Marshall studied and taught mathematics at Cambridge, before he became interested in economics. Thus there can be no doubt that he had a good knowledge of mathematics. He was distrustful of mathematical economics, because according to him, in real life variables at work are so numerous and interrelated that any attempt to put them into mathematical language would make the problem hopelessly complicated, and if one made the omissions required to make the problem manageable, this would yield an unrealistic construction. This, in essence, is the standard argument against mathematics, even today.

In the preface to his *Principles of Economics* which was the dominating textbook for more than 40 years (at least in England) and influenced generations of economists—he wrote in 1890: *his*

The chief use of pure mathematics in economic questions seems to be in

helping a person to write down quickly, shortly and exactly, some of his thoughts for his own use, and to make sure that he has enough, and only enough, premisses for his conclusions. But when a great many symbols have to be used, they become very laborious to anyone but the writer himself ... yet it seems doubtful whether anyone spends his time well in reading lengthy translations of economic doctrines into mathematics, that have not been made by himself (MARSHALL, 1961).

Statements of this sort made by this eminent economist had in my opinion serious consequences. Not only do I find them incorrect but they are not made with honest conviction. Only brilliance was demanded of a Victorian gentleman—not hard labour. In Samuelson's words, Marshall used mathematics to advance his inquiries 'in the boudoir', but not publicly. SCHUM-PETER (1941, p. 240) writes on this point, "...no serious objection can be raised to Marshall's acknowledgements to persons. But such objections is in order as regards his written and spoken comments about his impersonal ally to which he owed so much, mathematics." To substantiate the opinion that Marshall purposely played down the usefulness of mathematics, we should analyze his work in detail. It would then become quite clear that for a deep understanding of his arguments, one would need a great deal more of mathematics (and general equilibrium theory) than he admits to, in order to follow and understand his arguments. In a letter to Bowley in 1906 Marshall writes (MARSHALL, 1961, p. 775) that he follows these rules:

- "(1) *Use mathematics as a shorthand language, rather than an engine of inquiry.*
 - (2) *Keep to them till you have done.*
 - (3) *Translate into English.*
 - (4) *Then illustrate by examples that are important in real life.*
 - (5) *Burn the mathematics.*
 - (6) *If you can't succeed in (4), burn (3).*
- This last I did often!"*

Let us leave Marshall at this point, the case of Keynes is different. He was a man of vision. I agree with this statement of his: "Intuition will be in advance of analysis." But what does he really mean when he writes (KEYNES, 1936, pp. 297-298):

It is a great fault of symbolic pseudo-mathematical methods of formalizing a system of economic analysis, ... that they expressly assume strict independence between the factors involved and lose all their cogency and authority if this

hypothesis is disallowed; whereas, in ordinary discourse, where we are not blindly manipulating but know all the time what we are doing and what the words mean, we can keep ‘at the back of our heads’, the necessary reserves and qualifications and the adjustments which we shall have to make later on, in a way which we cannot keep complicated partial differentials ‘at the back’ of several pages of algebra which assume that they all vanish. Too large a proportion of recent ‘mathematical’ economics are mere concoctions, as imprecise as the initial assumptions they rest on, which allow the author to lose sight of the complexities and interdependences of the real world in a mass of pretentious and unhelpful symbols.

I can understand that to a man whose main interest is economic policy and who wants to make recommendation on urgent economic problems, logic must be a strait-jacket. Yet if we only knew what he kept ‘at the back of his head’, we could hopefully clarify the forty years of controversy about Keynes and Keynesians and what he really meant. There is no doubt, this book is full of contradictions—not really mistakes—since he changed his unspecified hypothesis from one chapter to the next.

So far the opinions of some great economists of the past. My own view is summarized in the following four statements:

- 1) Vision of facts and meanings precedes analytical work.
- 2) The quality of an economic theory does not depend on the depth of its mathematical content.
- 3) The rôle of mathematics in theoretical economics is essentially more than just a shorthand language or a convenient expository device.
- 4) The controversy about the rôle of mathematics is not one of principle, that is to say, mathematics yes or no, but a controversy about ‘how much’ and ‘what kind’ of mathematics—which amounts to saying that mathematics represents a psychological problem at any time to every economist.

Let me comment briefly upon these statements.

- 1) First of all I agree with Schumpeter’s thesis that, **in principle, vision of facts and meanings precedes analytical work.**

As a consequence of this thesis, mathematical methods are not helpful but rather misleading in those branches of economics where such a vision has not yet been formulated. Unfortunately, there are many branches of the social sciences which are still at this stage. Analytical work, and hence mathematics, presupposes a vision of facts and meanings sufficiently clear to allow the formulation of meaningful questions and hypotheses upon which

a theory can be built. Two famous examples of such visions are: Adam Smith's *The Wealth of Nations* and Keynes' *General Theory*.

Here I should mention that, as a matter of fact, visions of facts and meanings are in most cases based only on casual observations and intuitions but not on systematic statistical data analysis. For this reason, some econometricians are quite sceptical about economic theory in general. Their aim is not to give empirical content to economic theory but rather to analyze empirical data without relying upon any theoretical model. They consider theoretical models as being inadequate since quite often these models do not refer to the kind of data which is actually observed or, in principle, observable. This purely empiricist approach might have some advantages in special practical problems (short-run prognoses, for example), however, it has not been proved very successful. It is evident that statistical data analysis relies very much upon numerical mathematical methods.

2) The quality of an economic theory does not depend upon the depth of its mathematical content.

Economic theory is definitely more than, or should I perhaps say, different from a piece of mathematical reasoning. In this sense I agree with Marshall, who wrote in the previously referred to letter to Bowley that: "a good mathematical theorem dealing with economic hypothesis is very unlikely to be good economics".

This criterion of quality is different in mathematics than in economics. This obvious fact is too often forgotten, but typically less often by the skilled mathematical economist than by the mathematically inexperienced economist who is so easily impressed by symbols which he himself and his colleagues do not understand.

Quoting again Marshall, in his review of EDGEWORTH's *Mathematical Psychics* (1881) he expressed great fear that Edgeworth "might allow his mathematics to run away with him." Like quality this is a subjective judgement, and in the particular case Marshall was definitely wrong. But we should not concern ourselves here with the potential misuse of mathematics. In this respect, economics seems to be no special case. The potential and actual misuse of the literary style in economics seems to me much greater, but I do not want to pursue this point.

3) The rôle of mathematics in theoretical economics is essentially more than just a shorthand language or a convenient expository device.

Economic theory, if understood as an axiomatic theory, cannot be separated from mathematics. That economic theory should be formulated in an

axiomatic manner is not a recent requirement. Stanley JEVONS in his *Theory of political economy* (1871, p. 87) wrote in the chapter on "Logical Method of Economics" in 1871:

Possessing certain facts of observation, we frame a hypothesis as to the laws governing those facts, we reason from the hypothesis deductively to the results to be expected, and we then examine these results in connection with the facts in question.

There is a famous statement by David HILBERT in his *Axiomatisches Denken* (1918, p. 415):

Alles, was Gegenstand des wissenschaftlichen Denkens überhaupt sein kann, verfällt, sobald es zur Bildung einer Theorie reif ist, der axiomatischen Methode und damit mittelbar der Mathematik.

Contrary to the view often expressed in the literature, economists really have no choice between a purely literary and a mathematical presentation. Of course, hypothesis as well as conclusions can be formulated in prose. But, for example, the problem of the consistency of the hypothesis leads necessarily to a mathematical problem.²

I think, at this point, it is time to become a little more concrete. As a good example of an economic theory, let me describe in short the Walrasian Equilibrium Theory.

*Example: Walrasian equilibrium analysis.*³ I do not suggest that the Walrasian equilibrium theory or any other, say the Keynesian theory, is the most relevant model for our present day economic problems. I have chosen the Walrasian theory since it is particularly well suited for our purposes. I shall emphasize the conceptual, not the mathematical aspects of this theory.

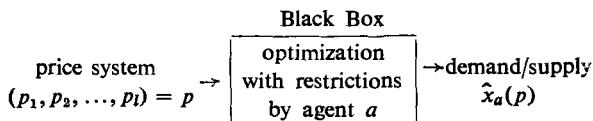
In this theory the economy consists of economic agents—consumers and producers—who make *decisions about commodities*; the amount of various commodities (consumption goods, labor, services, ...) they want to demand or supply at specified time periods. Thus, a consumer chooses a consumption plan (goods and services he plans to consume and type of labor he plans to supply) and a producer chooses a technologically feasible production plan (feasible input-output combination). Every agent is characterized by the limitations of his choice, and by his choice criterion.

Under what circumstances do the economic agents choose their consump-

² In saying this, I take for granted that we agree to call certain logical reasoning, mathematics. We all know that mathematics can be expressed verbally. Some people claim that mathematics is a language—fair enough—but if this is so, then it is a special one.

³ For a rigorous presentation of the Walrasian theory see DEBREU (1959).

tion or production plan? It is assumed that all agents face the same price system, which they consider as given and beyond their influence. It is further assumed that the consumption and production plans are the *result of an individual optimization problem*.



There are very simple but also quite sophisticated ways of specifying the agent's optimization problem. The various degrees of sophistication depend mainly on the more or less explicit treatment of time (the past, present, and future), assumptions about the availability of information and the modelling of other relevant factors for the individual decision process.

The economy is said to be *in equilibrium* if the price system \hat{p} is such that all individual plans $\hat{x}_a(\hat{p})$ are consistent, i.e., total demand equals total supply simultaneously on all markets. Observe that this definition of equilibrium does not rely on any process. This has often been criticized; equilibrium should be a stationary point of an explicitly defined price-commodity process. But there are great unsolved conceptual difficulties in specifying such a process in terms of economic agent's behavior.

Let me recall that the theory is designed to explain the rôle of the price system (a valuation of commodities) resulting from the interaction of the agents through markets for a competition economy in equilibrium.

The basic hypotheses are:

HYPOTHESIS 1 (on the behavior of individual agents): Economic agents believe that they can actually realize their plans—thus they believe in equilibrium. In forming their plans they act independently from each other, take the price system as given, and the plan, which is chosen, is the result of an optimization problem.

HYPOTHESIS 2 (on the behavior of market prices): The price system (whose rôle is to coordinate the independently taken individual decisions) is flexible and adjusts very quickly to a position where total demand equals total supply. Thus, any disequilibrium on a certain market is thought to be only temporary (transient).

Obviously, the two hypotheses are linked with each other. The question arises whether they are *consistent*. Now, if both hypotheses can be considered

as descriptive of actual behavior then the actual state of the economy, which is viewed as an equilibrium, may be considered as a proof of consistency.

This is clearly not a satisfactory answer. What, for example, is the meaning of the claim that the first hypothesis is descriptive? Can one test or falsify, in whatever sense, this hypothesis by observing actual behavior?

I do not want to elaborate this question here. Let me just mention that any observation which could contradict Hypothesis 1 could be interpreted as a situation where certain relevant but unobservable parameters, e.g., the preference relations which are underlying the optimization problem have changed.

If one holds the view that every hypothesis of an economic theory should in principle be directly falsifiable by observations, then, as a matter of fact, one has to reject most of economic theory.

There is an extensive literature with differing opinions on the methodological status of hypothesis in economics. The paper by Milton Friedman *The methodology of positive economics* (FRIEDMAN, 1935) and the replies to this paper, referred to and criticized by L. A. Boland in a recent paper (BOLAND, 1979) are good references.

Whatever position one takes in the controversy about the status of economic hypothesis, the request for a proof of logical consistence is legitimate. This question leads us, and there is no way of avoiding it, to a mathematical problem par excellence!

Indeed, even in the simplest case, it leads us to show that a system of equations has a solution, or equivalently, that a certain mapping has a fixed point.

Consider l commodities, and let $z_h(p)$ denote the total *excess demand* for commodity h if the price system $p = (p_1, \dots, p_l)$ prevails, (i.e., total demand minus total supply of commodity h). Then one has to show that the following system of equations has a solution.

$$\begin{aligned} z_1(p_1, \dots, p_l) &= 0, \\ \vdots &\quad \vdots \\ z_l(p_1, \dots, p_l) &= 0. \end{aligned}$$

It turns out that typically this system is not linear and has no special structure, say recursive. But obviously, it must have some properties in order to have a solution. In an axiomatic theory, these properties must be derived from assumptions on the characteristics of the economic agents, that is to say, those data which define the optimization problem $p \rightarrow \mathfrak{X}_a(p)$.

The simplest version of specifying the optimization problem is as follows:

A consumer is characterized by his *preferences* as for alternative consumption plans and his *wealth*. Preferences for consumer a are described mathematically by a binary relation \lesssim_a defined on the set X_a of a priori possible consumption plans. The wealth consists of two components; the endowment of material goods and the possessions of shares in profit.

In the simplest version of the theory the crucial simplification is that the individual preference relation \lesssim_a depends only on the past, which is considered as given, and hence is not treated explicitly. The preference relation \lesssim_a does not depend on current (or expected future) prices and consumption decisions of other agents (no externalities in consumption).

The *demand* $\mathfrak{x}_a(p)$ of consumer a is then defined as a maximal element with respect to the preference relation \lesssim_a subject to the budget restriction determined by the price system p and the wealth of consumer a .

A *producer* b is characterized by his *production set* Y_b which describes all technological feasible production plans. It is assumed that every producer b chooses a production plan $\mathfrak{x}_b(p)$ in Y_b which maximizes profit.

The theory is now sufficiently specified for asking the question of consistency, that is to say, which properties of the individual microeconomic concepts—preferences, wealth, and production sets—lead to an excess demand system of the economy which has a solution.

Since the final mathematical argument in the consistency proof will be the application of a fixed point theorem, we have to make sure that the excess demand $z(\cdot)$ is a continuous function in prices (or more general, a correspondence in prices with a closed graph). So, first of all, prices and quantities of commodities must be considered as real numbers (infinite divisibility) and one has to make sure that the individual decisions $\mathfrak{x}_a(p)$ depend continuously on the price system. This is easily achieved by assuming that the preference relations are continuous and the production sets are closed. These assumptions have no economic interpretation; as a mathematical consequence of the above idealization of commodities and prices, one has to make these assumptions.

But these continuity assumptions are clearly not sufficient. For example, we cannot conclude the existence of a maximal element $\mathfrak{x}_a(p)$, and, *a fortiori*, not its uniqueness.

The existence of a maximal element can be obtained by assuming that preference relations are either transitive or convex. For the uniqueness of a maximal element (which is not needed for the consistency proof, but simplifies it) one needs a much stronger assumption; the preference relation

has to be complete, transitive and strongly convex. This last property means that the convex combination of two consumption plans which are equivalent is preferred. These assumptions can be interpreted and are restrictive, in the sense that they restrict possible behavior. Many mathematical economists hesitate to make restrictive assumptions on individual preference relations, if this is done only to simplify the mathematical proof of consistency. After all, preferences are an economist's 'construct' in order to fill the 'black box'.

For this reason, for example, the convexity assumption of individual preferences has been replaced by the assumption that there are *many* consumers, in the sense that every individual consumer has only a negligible influence on collective actions. This leads to replacing a *finite set* of consumers by a *continuum* of consumers. This idealization, described mathematically by an atomless measure space of consumers, is more abstract, but this does not mean that it is less realistic (descriptive). This assumption implies that the excess demand $z(p)$ of the economy is a convex set (due to Liapunov's Theorem) which depends 'continuously' on the price vector p , and this is sufficient for applying a fixed point argument.

I shall not give, here, a complete list of assumptions on the micro economic concepts which guarantees the existence of an equilibrium. There is an extensive literature on this subject. In a recent survey paper by G. DEBREU, to appear in the *Handbook of Mathematical Economics*, one can find more than 250 references!

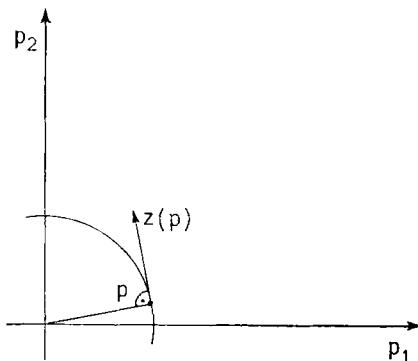
So far the existence of equilibria. Of course, the theory has not been designed just in order to show its consistency. There are two obvious questions economists are interested in. Firstly, does an equilibrium state of the economy have any efficiency or optimality properties? The answer is yes, in certain circumstances. To proceed in this direction would take up too much space.

Secondly, economists are interested in knowing how the equilibrium, in particular equilibrium prices, changes if certain data of the economy which have been considered as given, for example preferences or resources, vary. This is called 'comparative statics'. That conclusive comparative statics properties are desirable for applications of the theory goes without saying. But does the model in its present form have sufficient structure for yielding such results?

The equilibrium is unique only under very restrictive assumptions on the excess demand—which cannot be derived from assumptions on the micro concepts. Using methods of differential topology, one can show that

generically, equilibria are locally unique. But even locally, there are very few conclusive comparative statics results. Why is this so?

Let us again look at the excess demand systems $z(p) = 0$. From certain assumptions on the underlying micro data it follows that $z(\cdot)$ is a continuous function. Furthermore, one easily shows that $z(\cdot)$ is homogeneous of degree zero and fulfills the Walras identity, $p \cdot z(p) = 0$. Then we can normalize the price system so as to belong to the unit sphere and there we can interpret the excess demand function $z(\cdot)$ as a vector field.



An equilibrium corresponds to a singularity of the vector field. Does this vector field which is derived from an economy, have any special structure, for example, is it a gradient vector field? The answer is no. It has been shown by Sonnenschein and Debreu that for any given vector field one can construct an economy whose micro economic concepts have no pathological properties such that the excess demand derived from this economy coincides with the given vector field up to boundary behavior. This result sounds quite negative. Indeed, many propositions which economists thought to be derivable in this theory, cannot be expected. It is important to know this, even if it destroys illusions. This result shows the limits of the theory as sketched above. But the understanding of the limits of the theory might stimulate the development of a richer theory.

I hope that this short discussion of consistency and structure of the Walrasian equilibrium theory has illustrated that mathematics is more than just a shorthand language or expository device for theoretical economics.

4) In conclusion: the controversy about the rôle of mathematics is not one of principle, that is to say, mathematics yes or no, but a controversy about 'how much' and 'which kind' of mathematics.

Whatever position a person takes here, it is pretty weak, subjective and, as a matter of fact, highly correlated to his own actual acquired knowledge of mathematics.

Indeed, mathematics represents a psychological problem to every economist at any time.⁴

Edgeworth called calculus the 'mother-tongue of economics' since it was the appropriate method for the marginal utility school. But later, other problems could not be handled adequately with calculus. Koopmans pioneered with great success the use of convex analysis and made outstanding contributions. But recently when measure theory was introduced into economic analysis in order to formulate the traditional economic concept of perfect competition, he qualified (KOOPMANS, 1974, p. 327) the use of measure theoretic concepts as 'tours de force' and 'fanciful extensions'. Every economist has to face this psychological problem in his own way. Quite recently, when reading papers which employed Non-Standard Analysis in treating economic problems which I was accustomed to handle with standard measure theoretical tools, I felt this psychological problem very sharply. It is then in this situation so tempting and sounds so realistic to argue about the real world and to emphasize the weakness of the theoretical model in order to cover the lack of mathematical education.

References

- BOLAND, L. A., 1979, *A critique of Friedman's critics*, Journal of Economic Literature, vol. 17, pp. 503-522
- COURNOT, 1838, *Recherches sur les principes mathématiques de la théorie des richesses*, English translation by N. T. Bacon, Reprints of Economic Classics, 1960 (Augustus M. Kelley, New York)
- DEBREU, 1959, *Theory of value* (Yale University Press)
- EDGEWORTH, 1881, *Mathematical psychics* (C. Kegan Paul, London)
- FRIEDMAN, M., 1935, *The methodology of positive economics*, in: Essays in Positive Economics (University of Chicago Press)
- HILBERT, D., 1980, *Axiomatisches Denken*, Mathematische Annalen, vol. 78, pp. 405-425
- JEVONS, S., 1871, *The theory of political economy*, ed. R. D. Collison Black, 1970 (Penguin books)

⁴ A typical example is Novick's pamphlet (1954) and the replies by Samuelson, Klein, Dusenberry, Chipman, Tinbergen, Champerowne, Solow, Dorfman, and Koopmans.

- KEYNES, J. M., 1936, *The general theory of employment, interest and money, Collected writings of J. M. Keynes*, vol. 7, 1973 (MacMillan, London)
- KOOPMANS, 1974, *Is the theory of competitive equilibrium with it?* The American Economic Review, vol. 64, No. 2, pp. 325–329
- MARSHALL, A., 1961, *Principles of economics*, ed. C. W. Guilleband, vol. II, notes (MacMillan, London)
- NOVICK, 1954, *Mathematics: Logic, quantity and method*, Review of Economics and Statistics, Bd. 36, No. 4, pp. 357–358
- SCHUMPETER, J. A., 1941, *Alfred Marshall's principles* American Economic Review, vol. 31, No. 2, pp. 236–248
- SMITH, A., 1776, *An inquiry into the nature and causes of the wealth of nations* (London)
- WALRAS, L., 1900, *Éléments d'économie politique pure*, 4th ed., reprint 1976 (R. Pichon, Paris)

WORK OF PAUL BERNAYS AND KURT GÖDEL

GAISI TAKEUTI

University of Illinois, Urbana, Ill., U.S.A.

In the last two years, we lost two great logicians Paul Bernays and Kurt Gödel. Both Bernays and Gödel worked in proof theory and set theory and were deeply concerned over the foundation of mathematics.

Paul Bernays' logic career started in 1917, when Hilbert invited Bernays to Göttingen to work with him on the foundation of mathematics. From then till 1933, Bernays was Hilbert's coworker in Göttingen, the center of mathematics at the time. In 1904, Hilbert had proposed Hilbert's program to save mathematics from the crisis caused by contradictions in set theory. Hilbert's program consists of two parts.

1) *Formalization of mathematics.*

This is to make a mathematical theory a formal system in the following way.

First enumerate all primitive symbols in the theory. Then describe what kind of combination of symbols is a meaningful statement in the theory. Such kind of combination of symbols is called a formula. Finally describe what kind of sequence of formulas is a proof of the theory. A proof thus formalized is a concrete figure of symbols and is called a proof-figure.

2) *Consistency-proof of a formal theory.*

This is to provide a foundation for mathematics for proving its consistency. Here the meaning of consistency is the consistency of a formal theory, i.e. nonexistence of a proof-figure ending in a contradiction $0 = 1$. Since a proof-figure is a concrete figure of symbols, Hilbert thought that

his finite standpoint, i.e. a standpoint using only intuitive combinatorial arguments on concrete figure, is sufficient for the consistency proof. Hilbert's finite standpoint is more precisely the following: We can finitely operate on a concrete figure given before us, and infer a general statement as a Gedanken experiment.

The first part of Hilbert's program, i.e. the formalization of mathematics, had been already done at the time, e.g. in Whitehead and Russell's *Principia mathematica*. Therefore the second part is the actual Hilbert's program. For this purpose, one has to develop metamathematical theory of proof-figures based on the finite standpoint. Such a theory is called *proof-theory* or *Beweistheorie* in German. The joint work of Hilbert and Bernays of course belongs to proof-theory. Their work was written in detail in their two books *Grundlagen der Mathematik* (Springer, I, 1934 and II, 1939). The books were actually written line by line by Bernays. The main body of the books concerns the finite standpoint, the basic properties of the first order predicate calculus, ϵ -calculus, a proof-theoretic version of Gödel's completeness theorem, and the detailed proof of Gödel's incompleteness theorem. I will discuss Gödel's theorems later but it should be remarked here that Gödel only outlined his proof of the second part of his incompleteness theorem and its detailed proof was first carried out in Hilbert-Bernays' book.

First of all, I would like to say that the first chapter of volume 1 written by Bernays was the first coherent statement of what one might call Hilbert's program. As the most original part of the books, I would like to explain their theory of ϵ -calculus. In ϵ -calculus, we introduce ϵ -symbol in addition to the usual language of the first order predicate calculus. Then we add the following formation rule for terms. If $A(a)$ is a formula with a free variable a , then $\epsilon x A(x)$ is a term. The intended meaning of $\epsilon x A(x)$ is some x which satisfies $A(x)$ if there exists a such. This can be expressed formally only by the following axiom schema:

$$A(t) \rightarrow A(\epsilon x A(x))$$

where t is an arbitrary term. If one uses ϵ -symbol, he can eliminate quantifiers by replacing $\exists x A(x)$ by $A(\epsilon x A(x))$. On the other hand, if a theorem without ϵ -symbols is provable in ϵ -calculus, then a method is given to transform a proof-figure of the theorem in ϵ -calculus into a proof-figure without ϵ -symbol. They developed proof-theory on ϵ -calculus and proved Herbrand's theorem as an application of their theory. Today ϵ -calculus

is scarcely used. For example, I use Gentzen's sequential calculus in the place of ε -calculus. However, by looking at their books, I am convinced that Hilbert and Bernays developed the ε -method to such a degree that one can solve a problem on the first order predicate calculus by their theory as easily as by any other methods. In other words, they essentially completed the proof-theory of the first order predicate calculus.

In 1934, the political situation made Bernays go back to Zürich, where he taught 25 years in E.T.H. (The Eidgenössische Technische Hochschule). From 1937 to 1954 Bernays published seven papers on axiomatic set theory in the Journal of Symbolic Logic. A characteristic of this theory is that it has the class variables in addition to the set variables. As for his work on set theory, it is fair to say that he is the first person who organized set theory in the present standard and presented it in such a way that people could learn and use it very easily. For example, Gödel used Bernays' presentation of set theory in his famous monograph and many of us learned set theory from Gödel's monograph. In addition to the elegant presentation of set theory, Bernays showed which parts of axioms in set theory are used for what part of mathematics. For example, he introduced the axiom of dependent choice as the part of the axiom of choice necessary to analysis.

Evidently, a wide range of logicians profited from his work on set theory. In proof-theory, Bernays' influence is even greater. Kreisel's no-counter-example-interpretation is a continuation of the Hilbert–Bernays' ε -theorem. One cannot think of Feferman's work on arithmetization of metamathematics without Hilbert–Bernays' work of Gödel's theorem. Though Gentzen's teacher was Hermann Weyl, Gentzen's letters show that Bernays is his teacher in the true sense of being the one who encouraged him, advised him, and listened to him. Schütte, Prawitz, Maehara and myself among many others have worked in Gentzen's line. I would like to add one more important remark on Bernays' work. He made a great contribution to the philosophy of mathematics. Let me try to explain his philosophy of mathematics a little bit. He wrote "Mathematics, however, can be regarded as the theoretical phenomenology of structures." (Comments on *Ludwig Wittgenstein's Remarks on the Foundations of Mathematics*, Ratio II, No. I (1959), pp. 1–22). In his opinion, the main role of set theory is to provide us with models of structures. It seems to me that the superiority of Bernays' philosophy consists in avoiding the kind of simple-mindedness common among others. Bernays was the first to insist on the deceptiveness of the so-called conflicts between Brouwer's and Zermelo's or Cantor's views: when Brouwer spoke of contradictions with classical mathematics,

Bernays pointed out that the logical notions and the ranges of variables (choice sequences) had different meanings. So the ‘conflict’ consisted in different views of what is worth studying. Bernays also was the first to insist that Hilbert’s finitistic mathematics was a restriction of intuitionistic mathematics. His aphorism was “classical mathematics is the mathematics of Being, intuitionistic mathematics is the mathematics of Processes.” Dealing with the criticism of classical mathematics advanced by the intuitionists, he wrote “Heyting argues that it is too naive to ask questions like ‘Does there really exist a well-ordering of the continuum?’ But the question need not be put in this way. The question rather may be asked: ‘Is it suitable to adopt the strong extrapolation of classical analysis as it is formulated in the conceptions of set theory, in particular in the theory of transfinite cardinals?’ At the present time no deciding experience has been found to answer this question. Mathematicians have different opinions about the suitability of stronger or only weaker methods of idealizations. At all events the fact that our mathematical idealizations are so successful is a striking mental experience and by no means something trivial.” (*Mathematics as a domain of theoretical science and of mental experience*, Logic Colloquium ’73, pp. 1–4, Edited by H. E. Rose and J. C. Shepherdson, North Holland, 1975.) Bernays’ philosophy is much more subtle than the standard literature.

In 1930, Gödel proved the completeness theorem in his Ph.D. thesis which was later published as *Die Vollständigkeit der Axiome des logischen Funktionenkalküls*, Monatshefte für Mathematik und Physik, vol. 37, pp. 349–360, 1930. Previously in 1928, Hilbert and Ackermann formulated the first order predicate calculus in their book *Grundzüge der Theoretischen Logik* and stated: It is still an unsolved problem if the axiom system is at least complete in the sense that all logical formulas which hold in every structure, can be deduced. It can only be said empirically that this axiom system has sufficed in every application.

Gödel proved this affirmatively, i.e. that the first order predicate calculus is indeed complete. There is an interesting fact about this theorem as a historical event. In the eye of today’s logician, the completeness theorem is an immediate consequence of a lemma of Skolem in his 1922 paper: *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. But Skolem did not prove the completeness theorem and Gödel did. The completeness theorem can be expressed in

the following way: If a sentence is consistent, then it has a model. What Gödel proved was the following stronger form: If countably many sentences are consistent, then they have a countable model.

In 1931, Gödel proved the incompleteness theorem in his paper: *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme*, I, Monatshefte für Mathematik und Physik, vol. 38, pp. 173–198. There he considered the formal system of Principia Mathematica and proved:

- (1) If the system is ω -consistent, i.e. for no $A(x)$ are all of $A(0)$, $A(1)$, $A(2)$, ... and $\neg \forall x A(x)$ provable in the system, then there exists an arithmetical sentence φ such that either φ or $\neg \varphi$ is not provable in the system. (Rosser later improved the result by replacing ω -consistency by consistency.)
- (2) If the system is consistent, then there exists an arithmetical statement which expresses the consistency of the system and is not provable in the system.

Here one has to talk about the so-called *arithmetization of metamathematics* in order to explain how one finds an arithmetical statement expressing the consistency of the system. In the arithmetization of metamathematics, we first represent primitive symbols by natural numbers. Then formulas and proofs are represented by sequences of natural numbers and sequences of sequences of natural numbers. Since we can enumerate all the finite sequences of natural numbers, we can represent formulas and proofs by natural numbers called their *Gödel numbers* in such a way that different formulas or proofs correspond to different numbers. What Gödel did was to carry out an analysis showing that metamathematical properties like ‘ x is a formula’, ‘ x is a proof’, ‘ x is a provable formula’, etc. can be expressed by arithmetical formulas about the Gödel number x in the above correspondence. It follows that there exists an arithmetical statement which expresses the consistency of the system. It is because the system includes number theory that one can do many metamathematical deduction like mathematical induction on formalized metamathematical statements inside the system. Gödel obtained the above analysis in this way. By using it together with a diagonal argument, he succeeded in obtaining his results.

We can see here a typical Gödel work. Gödel used a formal system in two different ways. On one hand, he used it as merely rules on sequences of meaningless symbols. On the other hand, he used its mathematical meaning to carry metamathematics inside it.

After a little while, it was clear that his method went through on any natural axiomatic system including Peano's arithmetic.

Even today, Gödel's result is extremely impressive. It certainly gives us a deep insight on the nature of axiomatic system and consequently on our knowledge itself.

But it was even more sensational in the following reasons.

- (1) Hilbert together with many others took it for granted that there exists an axiomatic theory formalizing mathematics in such a way that for every arithmetical statement either the statement or its negation is provable in the system. Gödel's result showed that this belief is completely false.
- (2) Since the finite standpoint Hilbert originally conceived was rather elementary, it is obvious that any arguments from his finite standpoint can be formalized in Peano's arithmetic and therefore Hilbert's program, the central problem at the time, cannot be carried out from the finite standpoint Hilbert originally had in mind.

However, Gödel wrote at the end of his paper that his result does not rule out the possibility of finding some finitary consistency proof for mathematics in Hilbert's standpoint. Hilbert's standpoint merely requires the existence of a finitistic consistency-proof which cannot be represented in P (or in set theory or classical mathematics).

I do not know what Gödel meant by this remark. But I would like to add several comments here. In the finite standpoint—taken in a sense richer than Hilbert's—we can think of a concrete infinite sequence of concrete figures given before us instead of a single concrete figure and operate finitely on finite initial part of the sequence. Furthermore we can finitely operate on concrete operations, concrete operations on concrete operations, etc. and infer a general statement about them, as a *Gedanken experiment*. Actually in his 1936 paper: *Die Widerspruchsfreiheit der reinen Zahlentheorie*, Mathematische Annalen, vol. 112, pp. 493–565, Gentzen proved the consistency of Peano's arithmetic by using the accessibility up to the first ε -number and his proof of the accessibility can be justified in the finite standpoint described above.

Nevertheless Gödel's result made the original Hilbert's program cease to exist. However, two more remarks should be added here.

- (1) Gödel's notes left in the Institute for Advanced Study show that Gödel studied Gentzen's work very carefully. Actually, in his 1958

paper: *Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes*, *Dialectica*, vol. 12, #47/48, pp. 280–287, Gödel himself published a consistency-proof which is different from Gentzen's proof but deeply related to it.

- (2) From 1959 till 1974, I had many occasions to converse with Gödel. What he wished to discuss with me was mainly consistency-proof. I had an impression that he was seriously interested in the subject. Also he was the logician who encouraged my consistency work most.

In 1938, Gödel proved the consistency of the axiom of choice and the generalized continuum hypothesis with the axiom of set theory. (*The consistency of axiom of choice and the generalized continuum-hypothesis*, Proceeding of the National Academy of Science, vol. 24, pp. 556–557, 1938. *Consistency-proof for the generalized continuum hypothesis*, *Ibid*, vol. 25, pp. 220–224, 1939. *The consistency of the axiom of choice and of the generalized continuum-hypothesis with the axioms of set theory*, Annals of Mathematics Studies, no. 3, Princeton University Press, Princeton, NJ 1940.)

The result is very important. However, as Gödel himself believed (cf. S. C. Kleene: *An Addendum to "The Work of Kurt Gödel"*, *Journal of Symbolic Logic*, vol. 43, 1978, p. 613), his main achievement is his introduction of constructible sets. Gödel's notion of constructible sets comes from Whitehead and Russell's ramified hierarchy. Whitehead and Russell analyzed contradictions in set theory and found the vicious circle in the concept of sets as the source of contradictions. Thus they were lead to a concept of sets which is free from the vicious circle and very well justified philosophically. Their concept was called the *ramified hierarchy*. To explain it, let D be a well-defined set of well-defined objects like the set of all natural numbers. Let x, y, \dots range over the element of D . Since D is well defined, the notion $\forall x$ and $\exists y$ are also well-defined. Consequently, if $\varphi(x)$ is a formula of the first order predicate calculus starting with well-defined predicates on D like $<$ for natural numbers, then a set of the form $\{x|\varphi(x)\}$ is well-defined. Call these sets the *sets of ramified type 1* and introduce variables X^1, Y^1, \dots ranging over the set of ramified type 1. Now new quantifiers $\forall X^1$ and $\exists Y^1$ are also well-defined. Therefore the set of the form $\{x|\varphi(x)\}$ is well-defined if $\varphi(x)$ is obtained by adding $\forall X^1$ and $\exists Y^1$ to the first order language. Call these sets the *sets of ramified type 2*, and so on for every finite type n . Whitehead and Russell tried to construct classical analysis by using the ramified hierarchy. They could not do so and introduce the axiom of reducibility stating that any set of higher ramified

type is already among the sets of the first ramified type. But the introduction of the axiom reducibility changes the meaning of the ramified hierarchy and the original idea of the ramified hierarchy failed.

Gödel's construction of constructible sets is exactly the same as the construction of the ramified hierarchy except for the following two points.

(1) The construction is continued for all transfinite ordinals, i.e. ramified type α is constructed for every ordinal α . This can easily be done since Gödel's construction takes place inside set theory.

(2) In Gödel's construction, the domain D is also expanded as the construction goes one.

Gödel named the thus constructed sets as the *constructible sets* and denoted the class of all constructible sets by L .

Now the most interesting question here is whether all sets are constructible or not. Denoting the class of all sets by V , the question is expressed by whether $V = L$ holds or not. What Gödel proved is:

- (1) L is a model of set theory with $V = L$.
- (2) $V = L$ together with axioms of set theory implies the axiom of choice and the generalized continuum hypothesis.

Gödel's result shows that the continuum hypothesis is not a merely technical problem but a problem which is related with a fundamental question on set theory like "Is every set constructible?"

Modern set theory starts with this work of Gödel. This construction becomes a basic construction in set theory. For example, Paul Cohen proved the independence of the continuum hypothesis in 1963 by using a similar construction. Cohen carried out his construction in an imaginary universe while Gödel did it in the real universe.

As a historical remark, it should be mentioned that Hilbert proposed in his 1926 paper: *Über das Unendliche*, Mathematische Annalen, vol. 95, pp. 161–190, to prove the continuum hypothesis by iterating strictly constructive operations up to constructive ordinals. The combination of Whitehead and Russell's idea and Hilbert's idea is very close to Gödel's construction. Here again we can see a typical Gödel work. He used constructive idea in a nonconstructive way, whereas the others were either simply constructive or simply nonconstructive. In a note Gödel left in the Institute for Advanced Study, he explained that the Generalized Continuum Hypothesis in L is "nothing else but an axiom of reducibility for transfinite order" and carried out his construction and proof *à la* Hilbert.

I have discussed his three major works. I should do many other published works, e.g. a decidability result, his interpretation of classical logic in the intuitionistic logic, Gödel-Herbrand's definition of the recursive functions, his relativity work, and so on. But instead of these, let me talk about his 1947 paper: *What is Cantor's Continuum Problem?*, the American Mathematical Monthly, vol. 54, pp. 515–525. There he anticipated that the continuum hypothesis is independent from the present set theory and wrote “its undecidability from the axioms being assumed today can only mean that these axioms do not contain a complete description of the reality” and “the role of the continuum problem in set theory will be to lead to the discovery of new axioms which will make it possible to disprove Cantor's conjecture.” In his 1964 supplement of the same paper in *Philosophy of Mathematics* edited by P. Benacerraf and H. Putnam, Englewood Cliffs, NJ, Prentice Hall, pp. 269–273, he wrote further “despite their remoteness from sense experience, we do have something like a perception also of the object of set theory, as is seen from the fact that axioms force themselves upon us as being true.”

His opinion on set theory strongly influenced many set theorists. Search for new axioms especially axioms of strong infinity is now a substantial part of present set theory. The Continuum Problem seems to have been one of Gödel's biggest concerns. In the notes he left in the Institute for Advanced Study, we can see that he worked hard on many problems related to the continuum problem. It is no secret that he made an unsuccessful attempt in 1970 to prove $2^{\aleph_0} = \aleph_2$ by introducing new axioms.

At the end, I also would like to mention Gödel's unpublished works. Many notebooks he left in the Institute show that he was very active during 1940's and 1950's (more precisely till about 1955). Almost all his results of this period are not published. Unfortunately, I cannot discuss these works since they are written in Gabelsberger shorthand.

Gödel's results are basic in all major branches of logic today: model theory, recursive function theory, set theory, proof-theory, and intuitionism. Gödel is indeed the father of the present mathematical logic.

A SURVEY OF Π_2^1 -LOGIC

J.-Y. GIRARD

Sucy-en-Brie, France

The aim of Π_2^1 -logic is to play a role similar to ω -logic, but in a Π_2^1 context.

The general reason for the concepts of Π_2^1 -logic to play a great role in many problems comes from the contradictory tasks performed by the concepts:

(i) *effectivity*: these objects are indeed primitive recursive operations; the commutation to direct limits is indeed a way of using them as effective operations;

(ii) *logical complexity*: the concepts of ladder, β -proof are Π_2^1 -complete.

The basic patterns have been introduced in GIRARD (to appear); it is not possible to give a survey of this paper. However, we shall give some heuristics in Section 0.

The paper is essentially devoted to the exposition of Π_2^1 -logic through its main applications up to now (Aug. 79).

The heart of the paper is the cut-elimination theorem of Section 2, involving a cut-elimination procedure of a new kind.

Applications can roughly be divided into two classes:

(i) use of ladders for estimates connected with admissibles;

(ii) use of ladders to give (unusual) ordinal analysis.

The paper is supposed to be readable in the main lines by itself, but GIRARD (to appear) is absolutely necessary to a precise reading.

The reformulation of Jervell of the material of GIRARD (to appear) is of great interest to find examples, motivations (JERVELL).

Let us write a tableau comparing finite, ω , and β -logics:

Table 1. Comparison of finite, ω , and β -logics

| LOGICAL COMPLEXITY | Σ_1^0 | Π_1^1 | Π_2^1 |
|-----------------------------|----------------------|---|---|
| Typical objects | finite constructions | well-founded trees | functors commuting to $\lim \rightarrow$ |
| Completeness | finite proofs | ω -proofs | β -proofs |
| Interpolation | in the logic | in $L_{\omega, \omega}$ | in $L_{\beta\omega}$ (see VAUZEILLES) |
| Models | usual | ω -models | β -models |
| Cut elimination bound | exponential | ordinal exponential | Λ (in inductive logic) |
| Ordinal of PA | — | ε_0 | the Howard ordinal |
| Ordinal of ID ₁ | — | the Howard ordinal | the usual ordinal of ID ₂ |
| Finitistic reduction | Herbrand's theorem | no counter-example | Theorem 3.4 |
| Recursive hierarchies | — | Grzegorczyk | slow-growing |
| Intuitionistic completeness | yes | negative fragment (if $\exists K(K \rightarrow (\aleph_1)_{\omega}^{<\omega})$) (see VAUZEILLES) | negative fragment (if $\exists K(K \rightarrow (\aleph_1)_{\omega}^{<\omega})$) (see VAUZEILLES) |

0. An introduction to Π_2^1 -logic

In this section we shall give some milestones for the understanding of the concept of functor commuting to direct limits.

The basic category is the category ON of ordinals. The morphisms from x to y being given by the set $I(x, y)$ of strictly increasing functions from x into y .

0.1. EXAMPLE. Define for $x \in \text{ON}$ $F(x) = \omega^x$, and for $x, y, f, f \in I(x, y)$ define $F(f) \in I(\omega^x, \omega^y)$ by

$$F(f)(\omega^{a_0} n_0 + \dots + \omega^{a_p} n_p) = \omega^{f(a_0)} n_0 + \dots + \omega^{f(a_p)} n_p.$$

So F is a functor from ON into itself.

The basic property of F is the following:

Given any $x \in \text{ON}$ and $a \in F(x)$, there exist n and $f \in I(n, x)$ such that $a \in \text{rg } F(f)$.

(PROOF: If $a = \omega^{a_0}n_0 + \dots + \omega^{a_p}n_p$, let $n = p+1$, and let $f \in I(n, x)$ be: $f(0) = a_p, \dots, f(p) = a_0$; then $a = F(f)(\omega^p n_0 + \dots + \omega^0 n_p)$.)

This property of F is exactly the commutation to direct limits.

0.2. Limits and direct limits. Taking the supremum of a family of ordinals is a very natural way of approximation.

Everybody who has toyed with ordinal arithmetic knows that most of the operations are not continuous with respect to the supremum operation. For instance, $\omega + 1 \neq \sup(n+1)$.

However, the continuity can be restored via direct limits: We consider not only the integers $n+1$, but also the way each of them is embedded in the next one.

For instance, embed $n+1$ into $n+2$ by means of $f_n \in I(n+1, n+2)$:

$$\begin{aligned} f_n(x) &= x \quad \text{if } x < n, \\ f_n(n) &= n+1. \end{aligned}$$

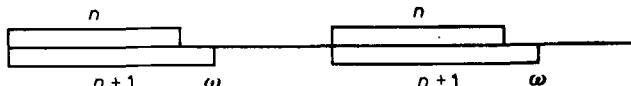
Then $\omega + 1$ naturally appears as the limit of this process. We write this

$$\omega + 1 = \lim_{\rightarrow}(n+1, f_n).$$

$\omega + 1$ is obtained as the union of the $n+1$, $n+1$ being identified with a subset of $n+2$ by means of f_n :



In fact, any ordinal is a direct limit of integers. In the case of a denumerable ordinal, the limit can be taken from a family (a_n) of integers with embeddings $f_n \in I(a_n, a_{n+1})$. For instance, in the case of $\omega + \omega$ with $a_n = n+n$ we have



0.3. Functors commuting to \lim . To say that F commutes to direct limits means that the image under F of a \lim is the \lim of the images. This gives an effective process of computing the values $F(\alpha)$: Suppose that we are given α under the form of a well-ordering of ω . Then α appears as a direct limit with $a_n = \text{order type of } \alpha \cap \{0, \dots, n-1\}$, f_n being the inclusion map between $\alpha \cap \{0, \dots, n-1\}$ and $\alpha \cap \{0, \dots, n\}$. Then $F(\alpha)$ will be computable by means of the system $(F(a_n), F(f_n))$.

Most of the functors commuting to \lim that I shall use are recursive (even primitive recursive). This means that the restriction of these functors to the category of finite ordinals $\text{ON} < \omega$ is indeed a functor from $\text{ON} < \omega$ into itself, and that this restricted functor is primitive recursive.

A functor from $\text{ON} < \omega$ into ON can always be extended to ON in order to commute to direct limits, but it may happen that the value $F(\alpha)$ is a linear order which is not a well-order. Such an extension is unique.

In fact, the property (for a primitive recursive functor from $\text{ON} < \omega$ into itself) of being extendable into a functor from ON into itself commuting to direct limits defines a Π_2^1 -complete set, as one may easily see from the β -completeness theorem of GIRARD (to appear).

The existence of extensions is the non-constructive part of Π_2^1 -logic.

If $\alpha \in S$ (where S is a cofinal subset of the first stable s_0 , see Section 4), then there exists a primitive recursive functor from $\text{ON} < \alpha$ into ON , commuting to \lim , which cannot be extended to the value $F(\alpha)$ (if α is infinite, of course). This can easily be obtained from the β -completeness theorem.

0.4. Are there many such functors? (i) The concept of such a functor is very general. This must be clear from the β -completeness theorem. However, it is obvious that there are ordinal functions which are not defined by functors, for instance a non-increasing function.

So the question is rather: what kind of functions can be expected to be majorizable by such a functor?

(ii) For instance, all ω_1^{CK} -recursive functions are bounded (for values α such that $\omega \leq \alpha < \omega_1^{\text{CK}}$) by a p.r. functor commuting to \lim . This result permits to reduce ω_1^{CK} -recursion to usual primitive recursion and (because of the majoration) to set theoretic primitive recursion.

Also, there is a functor Ω commuting to \lim such that for all $\alpha < \sigma_0$, $\Omega[\alpha] = \omega_\alpha^{\text{CK}}$. (This functor is not recursive at all; see 4.3.)

(iii) However, commutation to \lim gives a very obvious limitation of the growth.

* The set of all functors commuting to direct limits and sending integers into denumerable ordinals has the power of the continuum.

* Thus (using CH), one may construct a function from \aleph_1 into itself, growing faster than every functor commuting to \varinjlim .

(iv) It is important not to confuse commutation to \varinjlim and continuity of functions at limit points:

* Commutation to \varinjlim does not yield continuity. For instance, $F(x) = x + 1$, and, for all $f \in I(x, y)$, $F(f) \in I(x+1, y+1)$ defined by $F(f)(z) = z$ for $z < x$ and $F(f)(x) = y$ gives a functor commuting to \varinjlim , but $x \rightarrow x+1$ is not a continuous function.

* However, commutation to \varinjlim yields continuity when the following extra property holds:

$$F(E_{xy}) = E_{F(x)F(y)} \quad (E_{ab}(z) = z \text{ for } z < a; E_{ab} \in I(a, b)).$$

* Using CH, there are normal functions growing faster than any functor commuting to \varinjlim .

0.5. Functors and notational systems. Take any notational system we want, for instance Buchholz's system, (BUCHHOLZ, 1975). We shall work in the category $ON \leq \alpha_0$, where α_0 is the smallest denumerable ordinal such that $\bar{\theta}(\Omega_0, 0) = \bar{\theta}(\Omega_{\alpha_0+1}, 0)$. Let $F(x) = \bar{\theta}(\Omega_x, 0)$ and if $F \in I(x, y)$ define $F(f) \in I(F(x), F(y))$ by:

$F(f)(x)$ is obtained by replacing in the Buchholz notation for x all symbols Ω_i by $\Omega_{f(i)}$. General features of notational systems make $F(f)(x)$ be defined and $F(f)$ be strictly increasing.

This example clearly shows two things:

(i) *Functors commuting to direct limits can be introduced in every part of existing ordinal constructions.*

(ii) *Commutation to direct limits shows that the system up to $\bar{\theta}(\Omega_{\alpha_0}, 0)$ is already contained in its restriction to $\bar{\theta}(\Omega_\omega, 0)$.*

0.6. Ladders. Ladders are obtained from the concept of a rung: a *rung of type α* is something like a Bachmann collection of type α .

Given a rung R of type β and a function $f \in I(\alpha, \beta)$, one makes a rung ' R ' of type α as follows:

(i) first delete in R any interval $[x[a], x[a+1]]$ for $a \notin \text{rg}(f)$, and

(ii) collapse.

A *ladder* is a functor from ON into the category RG of rungs. This means that a ladder is a family $L(x)$ of rungs, $L(x)$ being of type x , and that if $f \in I(x, y)$, then $L(x) = {}^f L(y)$.

Of course, if one forgets the fundamental sequences, then a ladder induces a functor from ON into itself commuting to \lim .

0.7. β -*proofs*. Given any ordinal x , one can introduce the x -rule

$$\frac{A(0) \dots A(x') \dots \text{all } x' < x}{\forall z A(z)}$$

Now, if P is a proof with the y -rule (a y -proof) and if $f \in I(x, y)$, then one can try to construct a x -proof ${}^f P$:

- (i) delete in P all premises of some y -rule, of index $\notin \text{rg}(f)$;
- (ii) provided all remaining ordinal parameters are in $\text{rg}(f)$, then collapse.

A β -*proof* is a functor from ON into the category of proofs. This means that $P(x)$ is a x -proof and that if $f \in I(x, y)$, then

$$P(x) = {}^f P(y).$$

The tree structure of a β -proof can obviously be linearized in order to give a ladder.

0.8. *Global and local uses of the concept.* Let Φ be a function from ladders into themselves. The formula expressing that Φ maps ladders on ladders:

$$(\forall \alpha \in \text{ON } L(\alpha) \text{ is well-ordered}) \rightarrow (\forall \alpha \in \text{ON } (\Phi L)(\alpha) \text{ is well-ordered})$$

can be written

$$\forall \alpha \in \text{ON } \exists \beta \in \text{ON } \forall L \in L_\alpha (L(\alpha) \text{ is well-ordered} \rightarrow (\Phi L)(\alpha) \text{ is well-ordered})$$

or

$$\exists \tilde{\Phi} \forall \alpha \in \text{ON } \forall L \in L_\alpha (L(\tilde{\Phi}(\alpha)) \text{ is well-ordered} \rightarrow (\Phi L)(\alpha) \text{ is well-ordered}).$$

The interest of a construction on ladders depends on the $\tilde{\Phi}$. For instance, if $\tilde{\Phi}(\alpha) = \alpha$, then the construction (called *local*, because it does not really use the extension properties) is of no interest.

In the case of Λ (the one of GIRARD, to appear, as well as the one of Section 2), $\tilde{\Lambda}(\alpha) =$ the next admissible. Λ cannot be explained here—it is

something similar to the Bachmann hierarchy but in a very different context. May be, I shall give the reader the will of learning Λ if I say that Λ is certainly equivalent to some form of Bar-recursion of type 2.

1. Formalization of inductive definitions

We are starting with a language L ; given a positive operator $\Phi = \Phi(X, x)$ in L , one can form a new language $L' = L(D\Phi)$, where $D\Phi$ is a new unary predicate.

1.1. DEFINITION. Let T be a theory in the language L' and let \mathfrak{M} be a model of T .

(i) One defines the operator $\Phi_{\mathfrak{M}}$ from $\mathfrak{P}(|\mathfrak{M}|)$ into itself:

$$\begin{aligned} a \in \Phi_{\mathfrak{M}}(X) &\text{ iff } (\mathfrak{M}, X) \models \Phi(\bar{X}, x) \\ (X \subset |\mathfrak{M}|, \bar{X} &\text{ a new unary predicate letter}). \end{aligned}$$

(ii) \mathfrak{M} is said to be *inductive* iff

$$\mathfrak{M}(D\Phi) = \bigcap \{X \mid X \subset |\mathfrak{M}| \wedge \Phi_{\mathfrak{M}}(X) = X\}.$$

Our task is to characterize validity in all inductive models. Our main tool will be the β -completeness theorem of GIRARD (to appear). We are indeed looking for a bit more: a characterization with a deep proof-theoretic content.

1.2. DEFINITION. Let K be a ladder, and assume that $K = I + K' + 1$ for some ladder K' (I is the identity ladder, $1[\alpha] = 1_\alpha$ for all $\alpha \geq 1$).

Let k be the functor from (1) ON into itself associated with K . (So $k[\alpha] = \alpha + k'[\alpha] + 1$, $k[f] = f + k'[f] + E_1$. The reader unfamiliar with ladders can think of $k' = 0$ for this section.)

(i) For each $\alpha > 0$ we define $L'[K, \alpha]$ to be the first order language obtained from L by adding the unary predicates $D\Phi$ and $I\Phi^\lambda$, for all $\lambda < k[\alpha]$.

(ii) For each $\alpha > 0$, we define the concept of (K, α) -proof of a sequent $\Gamma \vdash A$ of $L'[K, \alpha]$, by the following inductive clauses:

(a) A set of axioms $P_1 \dots P_n \vdash Q_1 \dots Q_m$ where $P_1 \dots P_n, Q_1 \dots Q_m$ are atomic formulas of L ; we assume that this set is closed under cut, and that it contains the axioms $P \vdash P$ for P atomic in L .

(b) The usual logical rules of introduction and elimination for the connectives and quantifiers.

(c) The structural rules of weakening, exchange, contraction and cut.

(d) The logical rules for $D\Phi$ and the $I\Phi^1$:

$$\begin{array}{ccc} \text{Introductions} & & \text{Eliminations} \\ \frac{\Gamma \vdash A, \Phi(I\Phi^\lambda, t)}{\Gamma \vdash A, I\Phi^\mu(t)} \quad \lambda < \mu < k[\alpha] & \frac{\dots \Gamma, \Phi(I\Phi^\lambda, t) \vdash A \dots}{\Gamma, I\Phi^\mu(t) \vdash A} \quad \text{all } \lambda < \mu < k[\alpha] \\ \frac{\Gamma \vdash A, \Phi(I\Phi^\lambda, t)}{\Gamma \vdash A, D\Phi(t)} \quad \lambda < k[\alpha] & \frac{\dots \Gamma, \Phi(I\Phi^\lambda, t) \vdash A \dots}{\Gamma, D\Phi(t) \vdash A} \quad \text{all } \lambda < \alpha \end{array}$$

1.3. *Comments.* (i) The rules for $I\Phi^\mu$ clearly indicate the interpretation:

$$I\Phi^\mu = \bigcup_{\lambda < \mu} \Phi(I\Phi^\lambda).$$

(ii) The rules for $D\Phi$ are very unreasonable, because of the non-symmetry between the introduction and the elimination: in the introduction we allow λ to be arbitrary less than $k[\alpha]$ while we need only the proofs for all $\lambda < \alpha$ in order to perform the elimination.

The introduction can be read $I\Phi^{k[\alpha]} \subset D\Phi$ and the elimination can be read $D\Phi \subset I\Phi^\alpha$. Since $\Phi(I\Phi^\alpha) \subset I\Phi^{k[\alpha]}$, these rules express that $D\Phi = I\Phi^\alpha$ and that α is such that $I\Phi^\alpha = \Phi(I\Phi^\alpha)$. (Of course, $I\Phi^{k[\alpha]}$ used here is an abuse of notations.)

(iii) The sequents $A \vdash A$ are provable for all formulas A ; for instance:

$$\frac{\vdots}{\frac{\frac{\Phi(I\Phi^\lambda, t) \vdash \Phi(I\Phi^\lambda, t)}{\dots \Phi(I\Phi^\lambda, t) \vdash D\Phi(t) \dots \text{ all } \lambda < \alpha}}{D\Phi(t) \vdash D\Phi(t)}}$$

We leave the formal proof to the reader.

(iv) Let us show that the sequent

$$\Phi(D\Phi, x) \vdash D\Phi(x)$$

is (cut-free) provable:

$$\frac{\vdots}{\frac{\frac{\Phi(I\Phi^\lambda, t) \vdash \Phi(I\Phi^\lambda, t)}{\dots \Phi(I\Phi^\lambda, t) \vdash I\Phi^\alpha(t) \dots \text{ all } \lambda < \alpha}}{D\Phi(t) \vdash I\Phi^\alpha(t)}} \quad (1)$$

From (1) one easily constructs (using the fact that Φ is positive) a cut-free proof of

$$\Phi(D\Phi, x) \vdash \Phi(I\Phi^\alpha, x)$$

and we can infer by D -introduction the sequent

$$\Phi(D\Phi, x) \vdash D\Phi(x).$$

1.4. DEFINITION.

(i) If P_β is a (K, β) -proof and if $f \in (1) I(\alpha, \beta)$, then, by an obvious adaptation of the definitions in GIRARD (forthcoming), one can define (if it exists) a (K, α) -proof ${}^{k(f)}P_\alpha$.

(ii) We define the category \mathcal{C}_K as follows:

objects are 3-tuples $(\alpha, P, \Gamma \vdash A)$ where $\alpha \geq 1$, $\Gamma \vdash A$ is a sequent in $L'[K, \alpha]$ and P is a (K, α) -proof of $\Gamma \vdash A$;

a morphism from $(\alpha, P, \Gamma \vdash A)$ into $(\beta, P', \Gamma' \vdash A')$ is a function $f \in (1) I(\alpha, \beta)$ such that $\Gamma \vdash A = {}^{k(f)}(\Gamma' \vdash A')$, $P = {}^{k(f)}P'$.

(iii) A K -proof of the formula A of L' is a functor P from (1)-ON into \mathcal{C}_K such that

$$P[\alpha] = (\alpha, P_\alpha, \vdash A),$$

$$P[f] = f.$$

Such a functor obviously commutes to direct limits.

1.5. THEOREM. A is true in all inductive models (of axioms 1.2. (ii) (a)) iff it has a K -proof. One can even take $K = I+1$.

PROOF: (i) Let P be a K -proof of A , let \mathfrak{M} be an inductive model of the axioms, and let α be the closure ordinal of $\Phi_{\mathfrak{M}}$. Then we conclude (by an induction on P_α) that A is true in \mathfrak{M} . The fact that the rules of (K, α) -proofs preserve truth in \mathfrak{M} is obvious.

(ii) Conversely, if A is true in all inductive models, then one can easily show that A admits a functorial $(I+1)$ -proof under the extra axioms $\Phi(D\Phi, x) \vdash D\Phi(x)$ and in the D -elim rule we require $\lambda \leq \alpha$. Now one can delete the premises $\lambda = \alpha$ in the D -elims and since the extra axioms are provable, we are done.

The natural β -completeness involved here is Husson's theorem in HUSSON, applied to the generalized formulas of $L_{\alpha\alpha}$

$$I\Phi^\lambda(x) = \bigvee_{\lambda' < \lambda} \Phi(I\Phi^{\lambda'}, x) \quad (\lambda \leq \alpha),$$

$$D\Phi(x) = I\Phi^\alpha(x) \vee \Phi(I\Phi^\alpha, x).$$

2. The cut-elimination theorem

The non-symmetry between D -introd and D -elim makes cut-elimination very unreasonable. However, the cut-elimination theorem holds, but the price to pay for this is the tremendous increase of K .

I am sure that most of the readers will prefer to have a small hint of the proof, and go directly to the next section:

Suppose that we have a cut of the form

$$\frac{\begin{array}{c} \vdots^{P_\alpha} \\ \Gamma_\alpha \vdash A_\alpha, \Phi(I\Phi^\lambda, t)(\lambda < k[\alpha]) \dots \Gamma'_\alpha, \Phi(I\Phi^\mu, t) \vdash A'_\alpha \dots \text{all } \mu < \alpha \end{array}}{\Gamma_\alpha \vdash A_\alpha, D\Phi(t)} \quad \frac{\vdots^{Q_{\alpha, \mu}}}{\Gamma'_\alpha, D\Phi(t) \vdash A'_\alpha} \quad \frac{}{\Gamma, \Gamma' \vdash A, A'}$$

then the cut-elimination is delicate, because of the case $\lambda \geq \alpha$.

Now, observe that the proofs are indeed functors; and the cut can be removed as follows:

$$\frac{\vdots}{\Gamma_\alpha \vdash A_\alpha, \Phi(I\Phi^\lambda, t)} \quad \frac{\vdots^{Q_{k[\alpha]}, \lambda}}{\Gamma'_{k[\alpha]}, \Phi(I\Phi^\mu, t) \vdash A'_{k[\alpha]}} \quad \frac{}{\Gamma_\alpha, \Gamma'_{k[\alpha]} \vdash A_\alpha, A_{k[\alpha]}}$$

Observe that the obvious effect of the procedure is to replace k by $k \circ k$, and to change ordinal parameters; but only ordinals $\geq \alpha$ are changed, so we keep a reasonable control on this.

Now let us make the complete proof.

2.1. PROPOSITION. *If A is K -provable, then A is K -provable with the cut-rule restricted to unreasonable cuts*

$$\frac{\Gamma \vdash A, D\Phi(t) \quad \Gamma', D\Phi(t) \vdash A'}{\Gamma, \Gamma' \vdash A, A'}$$

PROOF: Easy adaptation (or consequence) of Husson's theorem (HUSSON), for the logic $L_{\beta\omega}$.

The proof of the main theorem is very similar to the construction of Λ (see GIRARD (forthcoming)).

Here V, V' will stand for admissible ordinals; if $\alpha < V$ and if K is a $[\alpha, V]$ or a $[\alpha, V]$ -ladder, then the concept of K -proof can easily be imagined by the reader. (Straightforward from 1.4. (iii).)

In general, one must consider the K -proofs of sequents $\Gamma \vdash A$ (which are of the form $\Gamma_\alpha \vdash A_\alpha$). If $\Gamma \vdash A$, we shall denote by $\tilde{\Gamma} \vdash \tilde{A}$ the result of replacing all $I\Phi^\lambda$ (with $\alpha \leq \lambda < k[\alpha]$) by $D\Phi$.

2.2. THEOREM. Let P be a K -proof of a sequent $\Gamma \vdash \Delta$ (K is an $[a, V]$ -ladder) with only unreasonable cuts. Then one defines an $[a, V]$ -ladder $\Lambda(P, K)$ and an $[a, V]$ -proof \hat{P} of $\hat{\Gamma} \vdash \hat{\Delta}$, with no cut formula containing $D\Phi$. The construction has the following properties.

(i) If $f \in (1) I(a', a)$, $K' = {}^f K$, $P' = {}^f P$, then

$$\Lambda(P', K') = {}^f \Lambda(P, K), \quad \hat{P}' = {}^f \hat{P}$$

(ii) if $a < V' \leq V$ and $K' = K \upharpoonright [a, V]$ -ON, $P' = P \upharpoonright [a, V]$ -ON, then

$$\Lambda(P', K') = \Lambda(P, K) \upharpoonright [a, V]$$
-ON and $\hat{P}' = \hat{P} \upharpoonright [a, V]$ -ON.

PROOF: This is done by induction on $\|P_V\|$. In fact, we define only the underlying ordinal functor of $\Lambda(P, K)$. (Because the ladder structure of K plays no role.) We do not verify (i) and (ii) which are trivial. In order not to abuse of the reader's patience we begin by precise treatments of the difficult cases, and we end by rough surveys of the trivial ones.

(1) The last rule of P_V is an (unreasonable) cut.

$$\frac{\begin{array}{c} :P_V^1 \\ \Gamma_V \vdash \Delta_V, D\Phi(t) \end{array} \quad \begin{array}{c} :P_V^2 \\ \Gamma'_V, D\Phi(t) \vdash \Delta'_V \end{array}}{\Gamma_V, \Gamma'_V \vdash \Delta_V \Delta'_V}$$

P_V^1 and P_V^2 being the values on V of K -proofs P^1 and P^2 .

(a) By induction hypothesis, we have obtained an L_1 -proof \hat{P}^1 of $\hat{\Gamma} \vdash \hat{\Delta}$, $D\Phi(t)$ and an L_2 -proof \hat{P}^2 of $\hat{\Gamma}', D\Phi(t) + \hat{\Delta}'$.

Let $L = L_2 L_1$.

(b) Construct an L -proof Q^1 of $\hat{\Gamma} \vdash \hat{\Delta}$, $I\Phi^{L_1[\alpha]}(t)$ by replacing in $\hat{P}^1[\alpha]$ every rule $\frac{\dots \Lambda \vdash \Pi, \Phi(I\Phi^\lambda, t)}{\Lambda \vdash \Pi, D\Phi(t)}$ introducing an ancestor of the cut-formula by the rule

$$\frac{\Delta' \vdash \Pi', \Phi(I\Phi^\lambda, t)}{\Delta' \vdash \Pi', I\Phi^{L_1[\alpha]}(t)}$$

(c) Construct an L -proof Q^2 of $\hat{\Gamma}', I\Phi^{L_1[\alpha]}(t) \vdash \hat{\Delta}'$ by replacing in $\hat{P}^2[l_1[\alpha]]$ every rule

$$\frac{\dots \Lambda, \Phi(I\Phi^\lambda, t) \vdash \Pi \dots}{\Lambda, D\Phi(t)} \quad \text{all } \lambda < l_1[\alpha]$$

eliminating an ancestor of the cut-formula by the rule

$$\frac{\dots \Lambda', \Phi(I\Phi^\lambda, t) \vdash \Pi' \dots}{\Lambda', I\Phi^{I^1[\alpha]}(t) \vdash \Pi'} \quad \text{all } \lambda < l_1[\alpha]$$

and by deleting all premises of index $\geq \alpha$ in the remaining D -elims.

(d)

$$\hat{P}[\alpha] = \frac{\begin{array}{c} Q^1[\alpha] \\ \hat{\Gamma} \vdash \hat{\Delta}, I\Phi^{l_1[\alpha]} \end{array}}{\begin{array}{c} Q^2[\alpha] \\ \hat{\Gamma}', I\Phi^{l_1[\alpha]} \vdash \hat{\Delta}' \\ \hat{\Gamma}, \hat{\Gamma}' \vdash \hat{\Delta}, \hat{\Delta}' \end{array}}$$

(2) The last rule of P_V is a D -elim or an I -elim with $\mu \geq \alpha$. Using the induction hypothesis, we obtain $[\sup(a, \lambda), V]$ ladders L_λ , and L_λ -proofs \hat{P}^λ , with \hat{P}_α^λ proving $\hat{\Gamma}$, $\Phi(I\Phi^\lambda, t) \vdash \hat{\Delta}$ for $a, \lambda \leq \alpha < V$.

There is an $[a, V]$ -ladder L such that

$$l(\alpha) = \alpha + \sum_{\lambda < \alpha} l'_\lambda(\alpha) + 1 \quad (\text{with } l_\lambda = i + l'_\lambda).$$

In the proof \hat{P}_α^λ replace each parameter $\alpha + \theta$ by $\alpha + \sum_{\lambda' < \lambda} l_{\lambda'}(\alpha) + \theta$. We obtain a new proof Q_α^λ .

Then define

$$\hat{P}[\alpha] = \frac{\dots \hat{\Gamma}, \Phi(I\Phi^\lambda, t) \vdash \hat{\Delta} \dots}{\hat{\Gamma}, D\Phi(t) \vdash \hat{\Delta}} \quad \text{all } \lambda < \alpha$$

(3) The last rule is a I -elim with $\lambda < \alpha$. Then one easily shows (because of the functorial character) that $\lambda < a$.

We do exactly as in (2) except that the conclusion is, of course, $\hat{\Gamma}, I\Phi^\lambda(t) \vdash \hat{\Delta}$.

(4) The last rule is a D -introd, or a I -introd with $\lambda \geq \alpha$. By induction hypothesis one has construct L' and \hat{P}' with \hat{P}'_α , a proof of $\hat{\Gamma} \vdash \hat{\Delta}$, $\Phi(D\Phi, t)$.

Let $L = L' + 1$ and replace all rules introducing an ancestor of one occurrence of $D\Phi$ in $\Phi(D\Phi, t)$:

$$\frac{\Lambda \vdash \Pi, \Phi(I\Phi^\lambda, u)}{\Lambda \vdash \Pi, D\Phi(u)}$$

by the rule

$$\frac{\Lambda' \vdash \Pi', \Phi(I\Phi^\lambda, u)}{\Lambda' \vdash \Pi', I\Phi^{I^1[\alpha]}(u)}.$$

So we obtain an L' -proof Q of $\hat{\Gamma} \vdash \hat{A}, \Phi(I\Phi'^{[\alpha]}, u)$, and we define

$$\hat{P}[\alpha] = \frac{Q_\alpha}{\hat{\Gamma} \vdash \hat{A}, D\Phi(u)}.$$

(5) All the remaining cases are obvious. In the case of two premise rules, one must, given a L_1 -proof and a L_2 -proof, change them into an L -proof, with the same L . $L = L_1 + L_2$ or $L = L_2 L_1$ do it very well...

2.3. PROPOSITION. *Given a K -proof with only cuts on formulas not containing $D\Phi$, there exists a cut-free K -proof of the same formula.*

PROOF: Trivial from Husson's theorem.

2.4. THEOREM. *Suppose that A is K -provable for a certain ladder K ; then A is cut-free K' -provable for a certain ladder K' .*

PROOF: Obvious from 2.1, 2.2, 2.3.

3. Translation of the cut-elimination theorem

We shall translate the cut-elimination theorem in non-proof theoretic terms. The basis for such a translation is given in GIRARD (1976), part A.

3.1. DEFINITION. Let A be a formula of L' , and K be a ladder. We define the (functorial) formulas $?A_K, !A_K$ by

$$?A_K[\alpha] = ?A_{K,\alpha}, \quad !A_K[\alpha] = !A_{K,\alpha},$$

$?A_{K,\alpha}$ (resp. $!A_{K,\alpha}$) being obtained from A by replacing every positive (resp. negative) occurrence of $D\Phi$ by $I\Phi^{k[\alpha]}$ and every negative (resp. positive) occurrence of $D\Phi$ by $I\Phi^\alpha$.

3.2 THEOREM. *The following are equivalent:*

- (i) A is true in all inductive models.
- (ii) A is I+1-provable.
- (iii) A is K -provable without cuts for some recursive ladder K .
- (iv) $?A_{K,\alpha}$ is true for all α , for some recursive ladder K .
- (v) $?A_K$ is β -provable, for some recursive ladder K .

3.3. Remarks. (i) Write $A = \mathcal{A}[D\Phi^-, D\Phi^+]$ by separating negative and positive occurrences of $D\Phi$. The validity of A means that $\mathcal{A}[I\Phi^{\alpha_0}, I\Phi^{\alpha_0}]$, where α_0 is the closure ordinal of Φ . From this it follows that $\mathcal{A}[I\Phi^\alpha, I\Phi^{\alpha_0}]$ is valid for all α .

Now Theorem 3.2 permits to improve the value α_0 found for the positive occurrences, because $?A_{k,\alpha} = \mathcal{A}[I\Phi^\alpha, I\Phi^{k[\alpha]}]$.

(ii) The cut-rule appears here as the general principle $?A = !A$ (see GIRARD, 1976, part A). This principle says exactly that α is the closure ordinal of Φ . The cut-elimination theorem allows us to eliminate this principle. This principle is of course the impredicative part of inductive definitions.

The cut-elimination theorem eliminates thus the impredicative axioms; but since there is no miracle in that kind of problems, the impredicativity has been transferred to K : the set of recursive ladders in Π_2^1 -complete.

3.4. THEOREM. In the language of arithmetic, consider the following operator

$$N(X, x): x = \bar{0} \vee \exists y (y \in X \wedge x = Sy).$$

Then the interpretation of IN^α is the set of integers $<\alpha$, while DN is the set of integers.

Let $A = \forall x \exists y \forall z \exists t R(x, y, z, t)$ with R primitive recursive be a formula of arithmetic; then the following are equivalent:

- (i) A is true;
- (ii) there is a recursive ladder K such that

$$\forall x \in IN^\alpha \exists y \in IN^{k[\alpha]} \forall z \in IN^\alpha \exists t \in IN^{k[\alpha]} R(x, y, z, t)$$

is true for all α ;

- (iii) for some recursive ladder K , the formula of (ii) is β -provable.

3.5. Remarks. (i) The formulation 3.4 does not need the fact that A is prenex.

(ii) Kreisel's no-counterexample interpretation (KREISEL) gives a characterization of truth by means of continuous functionals.

The no-counterexample relies heavily on $\Pi_1^1 (= \omega)$ -logic. This theorem is in some sense the Π_2^1 analogous of Kreisel's theorem.

3.6. THEOREM. In the language L of arithmetic + DN (see 3.4) one defines

$$O(X, x): x = 1 \vee \exists y (y \in X \wedge x = 2^y)$$

$$\vee \exists y (x = 3 \cdot 5^y \wedge \forall z \in DN \exists w (T_1(y, z, w) \wedge U(w) \in X \wedge w \in DN)).$$

Suppose that a formula $\forall x \in DO \exists y \in DO R(x, y)$ (R arithmetical) is true. Then there exists a primitive recursive K such that, for all $\alpha \geq \omega$,

$$\forall x \in IO^\alpha \exists y \in IO^{k[\alpha]} R(x, y).$$

PROOF: One must apply a straightforward extension of the cut-elimination theorem. Observe that $!O_{K,\alpha} = IO^\alpha$ and $?O_{K,\alpha} = IO^{k[\alpha]}$ for $\alpha \geq \omega$ (because $IN^\alpha = IN^{k[\alpha]} = DN$ in that case). In order to conclude our proof we need only to show that any recursive ladder is bounded (for infinite values) by a primitive recursive ladder—this is left to the reader.

3.7. COROLLARY. *Every function Σ^1 on $L_{\omega_1 \cdot k}$ is bounded (for values $\geq \omega$) by some $k[\alpha]$, where K is a primitive recursive ladder.*

This result has also been proved by MASSERON by direct methods.

4. Applications to admissible set theory

One can generalize 3.6 and 3.7 to a relatively large class of admissibles.

4.1. DEFINITION. (i) α is said to be *weakly stable* iff all $[1, \alpha]$ -ladders L such that $L(\omega) \in L_\alpha$ are such that one can extend them into $[1, \alpha]$ -ladders, i.e. the value $L(\alpha)$ computed by means of direct limits is well-founded. We denote by σ_0 the first weakly stable.

(ii) We define the set S of ordinals by $\alpha \in S$ iff there exists a primitive recursive $[1, \alpha]$ -ladder not extendable into a $[1, \alpha]$ -ladder.

4.2. Remarks. (i) Obviously, σ_0 is the first ordinal not in S .

(ii) Every stable ordinal (i.e. such that L_α is a Σ^1 -elementary substructure of L) is weakly stable; but σ_0 is not stable.

(iii) Using the Π_2^1 -completeness of ladders, it is easy to show that $S \subset s_0$ (where s_0 is the first stable) and that S is cofinal in s_0 .

4.3. PROPOSITION. *The following are equivalent:*

(i) $\alpha \in S$.

(ii) *There exists a primitive recursive β -theory T which is β -consistent and such that all β -models M of T are such that $M(ON) = \alpha$.*

4.4. THEOREM. *Let $\alpha < \sigma_0$, and let α^+ be the next admissible; then*

(i) *every function Σ^1 on L_{α^+} is bounded for values $\geq \alpha$ by $k[\cdot]$, where K is a primitive recursive ladder;*

$$(ii) \alpha^+ = \sup_{K \text{ primitive recursive}} k[\alpha] = \sum_{K \text{ primitive recursive}} k[\alpha] = l[\alpha]$$

for a certain (non-recursive) functor l from (1) ON into itself, commuting to \lim .

PROOF: (i) Let T be a theory as in 4.3(ii); one can assume that L_α is definable in T ; now, L_{α^+} can be obtained from T by an inductive definition... and it is possible to conclude as in 3.6, 3.7.

(ii) is obvious: l is defined by the last equation.

Remark. If $\alpha \notin S$, then $l(\alpha) \leq j(\alpha) < \alpha^+$, where j is the sum of all k , where K varies through all primitive recursive $[1, \alpha]$ -ladders. The fact that j is defined is a consequence of $\alpha \notin S$.

4.6. THEOREM. *There exists a functor Ω from (1) ON into itself, commuting to direct limits and for all $\alpha < \sigma_0$, we have $\Omega(\alpha) = \omega_\alpha^{\text{CK}}$.*

PROOF: Ω is easily obtained from l (left to the reader).

5. The proof theoretic significance of these results

First we state a theorem more or less natural from the literature. This result expresses the essence of the Gentzen-Schütte proof theory.

The system PRA² is a second-order logic with comprehension and induction restricted to formulas whose only quantifiers are bounded first-order quantifiers; the language is supposed to include symbols for the primitive recursive functions.

5.1. THEOREM. *In PRA², the following are formally equivalent:*

- (i) *for all X the jump of X exists;*
- (ii) Π_1^0 —CA;
- (iii) *in any system of ω -logic with axioms of complexity n , if A is provable with cuts of complexity $n+1$, then A is provable with cuts of complexity n ;*
- (iv) *if α is well-ordered, so is 2^α .*

PROOF: (i) \leftrightarrow (ii) is trivial.

(ii) implies (iv). Let $f(n) = 2^{a_0^n} + \dots + 2^{a_n^n}$ be a s.d.s. in 2^α , and assume that α is well-ordered. Then define N_i as follows:

- $a_0^n \geq a_0^{n+1}$, so, since α is well-ordered, $\exists N_0$ such that

$$a_0^p = a_0^{N_0} \quad \text{for all } p \geq N_0,$$

• then for all $p > N_0$ we have $i_n \geq 1$, so $2^{a_i^n}$ is decreasing.

Choose $N_1 > N_0$ such that $a_1^p = a_1^{N_1}$ for all $p \geq N_1$...

• The sequence $a_0^{N_0}, a_1^{N_1}, a_2^{N_2} \dots$ is a s.d.s. in α .

(iv) implies (iii), because of the estimates on the height of the proof-tree arising from cut-elimination.

(iii) implies (ii). Given a set of pairs of integers $X(n, m)$, one may take as axioms all true sequents

$$X(\bar{n}_1, \bar{m}_1), \dots, X(\bar{n}_p, \bar{m}_p) \vdash X(\bar{n}_{p+1}, \bar{m}_{p+1}), \dots, X(\bar{n}_q, \bar{m}_q).$$

Now one easily constructs a (may be not well-founded) ω -proof of the void sequent \vdash , this proof being primitive recursive in X . The idea is to introduce for all p, I, J, f such that $I \cap J = \emptyset, I \cup J = p$, and f is a function from J to ω , a sequent S_{IJf} :

$$\dots \forall x X(i, x), X(\bar{i}, \bar{0}), \dots, X(\bar{i}, \overline{p-1}), \dots \vdash \dots$$

$$\dots \forall x X(j, x) X(\bar{j}, \overline{f(j)}), \dots (i \in I, j \in J)$$

Such a sequent is said to be *past secured* when $S_{I'J'f'}$ is a weakening of an axiom, where $I' = I \cap p-1, J' = J \cap p-1$ and $f' = f \upharpoonright J'$ ($p > 0$). The proof-tree consists essentially of the non-past-secured S_{IJf} ; if S_{IJf} is not a weakening of an axiom, then it can be established from $S_{I \cup \{p\}Jf}$ and all the $S_{I \cup \{p\}g}$ by means of the rules for ω , and a cut on $\forall x X(\bar{p}, x) \dots$

A s.d.s. in this proof would give the set $\{p \mid \forall x X(\bar{p}, x)\}$ primitive recursively in the sequence. So, if this set does not exist, then the proof is well-founded; if the proof is well-founded, then by (iii) one can eliminate cuts of degree 1. But a proof of \vdash with only cuts of degree 0 is necessarily finite, and is easily shown not to exist.

We have the more or less perfect analogous for inductive logic:

5.2. THEOREM. In PRA² the following are equivalent:

- (i) for all X the hyperjump of X exists;
- (ii) Π_1^1 -CA;
- (iii) in any system of inductive logic, one can eliminate the cuts;
- (iv) if L is a ladder, then ΛL is a ladder.

PROOF: (i) and (ii) are obviously equivalent.

(ii) \rightarrow (iv): see GIRARD (forthcoming).

(iv) \rightarrow (iii): This is indeed an annoying thing; one would have to compare the Λ of GIRARD (forthcoming) with the one of Theorem 2.2. This presents no interest and no difficulty.

(iii) \rightarrow (ii): Let Y be a set of integers. Then we form a theory T^Y exactly as in 3.6 unless that

- add the true sequents $\bar{n}_1 \in Y, \dots, \bar{n}_p \in Y \vdash \bar{m}_1 \in Y, \dots, \bar{m}_q \in Y$;
- replace T_1 by T_1^Y .

Then the cut-elimination theorem shows that the system is consistent. So it has an inductive model, which is the hyperjump of Y . The implication consistent \rightarrow inductive model is easily provable from Π_1^0 -CA. The reader will easily show (using 5.1) that (iii) $\rightarrow \Pi_1^0$ -CA.

5.3. THEOREM. *Suppose that a Π_2^0 -formula is K-provable in a system of inductive logic, cut-free. Then the recursive Skolem function of the formula is bounded by*

$$k[n] = \gamma_{K[\omega]}(n)$$

where γ is the slow-growing hierarchy:

$$\begin{aligned} \gamma_0(n) &= 0, \\ \gamma_{\alpha+1}(n) &= \gamma_\alpha(n)+1, \\ \gamma_\lambda(n) &= \gamma_{\lambda[n]}(n) \quad \text{for } |\lambda| \text{ limit.} \end{aligned}$$

5.4. General significance of 5.3. (i) The relation $\lambda_{L(\omega)}(m) = \gamma_{(AL)(\omega)}(m)$ (GIRARD, forthcoming) opened a very upsetting problem. The hierarchy λ was connected with the systems $ID_0 (= PA) ID_1 \dots ID_n$ by means of the usual proof-theoretic ordinals $|ID_n|$ of the theories:

f is provably recursive in $ID_n \rightarrow f$ bounded by $\lambda_{|ID_n|}$.

But the connection, by means of γ , yielded the ordinal $|ID_{n+1}|$ for the theory ID_n . So the problem was to find a natural proof-theoretic reason for this.

(ii) This is achieved in the present paper. The cut-elimination theorem yields a natural cut-elimination procedure for theories of inductive definitions, which gives the odd ordinal numbers as the values $k[\omega]$.

(iii) So, in the case of arithmetic, we have a cut-elimination procedure that yields the Howard ordinal $|\Lambda \varepsilon_0(\omega)|$.

Such a procedure is of course absolutely independent of the familiar Gentzen-Schütte procedures.

5.5. EXERCISES. (i) Cut elimination for the Π_1^1 -comprehension axiom, using the finite iteration of the procedure $\Phi \rightarrow D\Phi$, for formulas with no second order quantifiers, and comprehension for these formulas. Comparison with Takeuti's proof (TAKEUTI, 1967).

(ii) Cut elimination for the ID's, by use of (v) ladders, the process for v being a functor of v commuting to \lim . Comparison with Pohlers' method (POHLERS, to appear).

Added in proof (October 1980): The paper Π^1_2 -logic of GIRARD is not yet appeared, but its contents will be quite different from what was expected in August 79. This implies the following simplifications in the paper: it is no longer necessary to consider ladders, because dilators (functors from ON to ON commuting to direct limits and to pull-backs) do exactly the same job. Concretely, this means that one can identify in this paper, a ladder with its underlying ordinal functor. The cut-elimination theorem of Section 2 has been proved in a more detailed version in a paper of GIRARD, to appear in the volume of "l'Enseignement Mathématique", dedicated to E. Specker (proof theoretic investigations of inductive definitions (part 1)). Results of Section 4 have been improved by Jacques Van de Wiele, in his *thèse de troisième cycle*: he considers ∞ 0-recursive functions.

References

- BUCHHOLZ, W., 1975, *Normalfunktionen und Konstruktive Systeme von Ordinalzahlen*, Lecture Notes 500 (Springer-Verlag)
- GIRARD, J. Y., 1976, *Three-valued logic and cut-elimination*, Dissertationes Mathematicae, vol. 136
- GIRARD, J. Y., Π^1_2 -Logic, Part 1, Annals of Mathematical Logic (to appear)
- HUSSON, J. F., *Complétude et élimination des coupures pour $L_{\beta\omega}$* , Thèse de troisième cycle Université Paris VII (Paris)
- JERVELL, H. R., *Homogeneous trees* (Underground)
- KREISEL, G., *On the interpretation of non-finitist proofs II*, the Journal of Symbolic Logic, vol. 17
- MASSERON, M., *Majoration des fonctions ω_1^{CK} -récursives par des échelles*, Thèse de troisième cycle Université Paris X (Paris)
- POHLERS, W., *Cut-elimination for impredicative systems*, (Part 1-Part 2), Archiv für mathematische Logik (to appear)
- TAKEUTI, G., 1967, *Consistency proofs of subsystems of classical analysis*, Annals of Mathematics, vol. 86
- VAUZEILLES, J., *Complétude et interpolation en β -logique*, Thèse de troisième cycle Université Paris VII (Paris)

LOGICAL APPROACH TO PROGRAMMING

N. N. NEPEIVODA

1. Preliminaries

The basic philosophical background to the logical approach to programming (LAP) consists of the following

- (a) The process of program construction is considered to be a search for a proof in a corresponding formal theory.
- (b) We are concerned mainly with composing a program from given subroutines. Thus the basic notions of computability are the ones of abstract computability.
- (c) The corresponding formal theories and logics may be different at different stages of a proof. (For example, to construct an interactive system we may use action relevancy logic for transput; intuitionistic logic for routines; classical logic for convergency proofs.)
- (d) Proofs and programs join together under the primacy of proof. A program is the executable part of a proof.
- (e) We sufficiently stress simple combinations of well-known methods and principles (combinational problems). Correspondingly we concentrate on weak notions of computability; we will not use full proof-searching methods.
- (f) The best traditions of mathematical logic and computer programming are critically appreciated. For example, the yeath of ideas of mathematical logic are expressed in an overcomplicated and inconvenient form. Well designed algorithmic languages force us to represent our ideas in an inadequate form. We must write precisely those things, that can be reconstructed automatically, if we have an adequate representation of our ideas. We must omit the ideas that are necessary to explain our program.

The origin of LAP was in the speach of A. A. Markov in October 1968 at Moscow University. Markov set the problem of applying constructive

mathematics to programming. He described the main mistakes of theoretical programming (it considers a program independently of its construction) and outlined possible paths for research. Some years later (in 1971) similar reasons were given in the work of R. L. CONSTABLE (1971). The author begun to work in this area in 1977. The work of KREISEL (1975) was very helpful in advancing the investigations.

Essentially LAP is *the applied theory of proofs*. We must join strong proof-theoretic methods and programmatic design together. From very beginning LAP posed many problems. The main ones are the following.

- (i) To single out natural classes of formal theories for which there exist simple algorithms for extracting programs from proofs.
- (ii) To design algorithmic languages whose structures reflects the structure of proofs adequately.
- (iii) To give a new technique of formalization, because formal theories are to be regarded as metaprograms.
- (iv) To give a formalization of the notion of proof convenient for expressing the ideas of proofs. Design of this must be comparable with that of modern algorithmic languages.
- (v) To separate and to formalize new logics which arise in different branches of programming.
- (vi) To give convenient methods of composing proofs constructed by means of different logics into a single one.
- (vii) The resolution method and its generalizations are inconvenient in this area. We need a new system of automatic demonstrations.
- (viii) To give adequate semantics for the program development.

We look at logics of programming rather than at program logic. LAP has allowed us (NEPEIVODA, 1978a, b, 1979a, b, c, d, KOGAN, 1978)

- to set precisely some old problems of programming;
- to give some criteria of valuation of programming languages;
- to reexamine the valuability of some traditional branches of theoretical computer science;
- to work out the natural method of teaching the art of programming.

2. Three approaches to program synthesis

In 1978 we had two different approaches to the problem of synthesis. Taking the transformational approach (ERSHOV, 1977) we start from a problem description and a language which is simultaneously descriptive

and ‘theoretically’ algorithmic. This description gives us a (possibly very inefficient) algorithm. We must now transfer from *computability* to *computation*. To do this we apply various rules of program transformation. As soon as the quality of a transformed program becomes good enough we stop.

The approach of Manna (MANNA, 1977) is completely different. With this approach we formulate a problem by means of a descriptive language. Then we wrench a piece of this problem out and try to synthesize a program realizing this piece. Afterwards we transform this piece of the program in such a way that it satisfies a new wrenched piece of the problem and the old one simultaneously. This approach corresponds to the ‘bottom-up’ program development.

We can see the cyclic character of the approaches considered. LAP also has its own cycle of program development. We set a question by asserting a logical statement. After proving it we can obtain a theoretically correct program, but its efficiency cannot be taken for granted. Moreover, the program constructed can be unrealizable by the given computer. Thus we must analyze the program, extract new requirements and try to transform it in order to satisfy them. If our transforming becomes useless we must find a new ‘theoretical’ solution counting with the new specifications. We see that Ershov’s approach is included naturally in LAP.

We can illustrate these differences by means of the three figures:

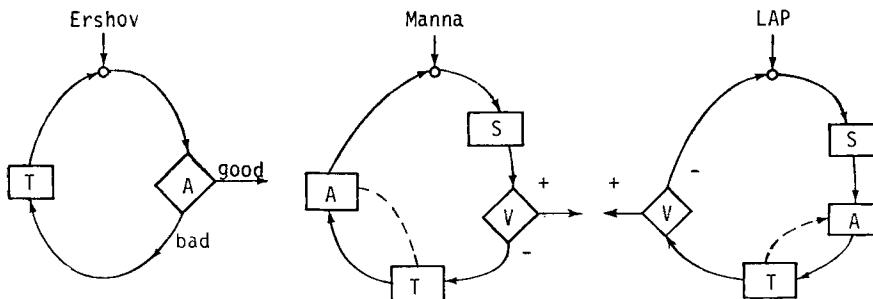


Fig. 1

S—synthesis, A—analysis, T—transformation, V—verification.

We point out that analysis and verification can be informal in LAP.

Let us compare these three approaches. Ershov’s approach is now the unique one applicable to complex scientific programs. Manna’s one is the

most stable relative to change of description and change of algorithmic languages. Ershov's approach demands a well connected family of algorithmic languages. Good and coordinated design of logical and algorithmic languages is demanded by LAP. From another side, there is some kind of slavery in Manna's approach. This approach does not try to change traditional mistakes in programming language design. The synthesis adapts itself to the existing circumstances. Sometimes (for example for conditionals) Manna's approach seems to be completely ill-considered. But in the case of ill-considered problems not precisely formulated this approach is the best one.

3. Pure programming

Let us first consider the traditional case. We must construct the method of computation of the output value y for each x satisfying the input condition $A(x)$. The input-output relation is $B(x, y)$ and this can be expressed by

$$\forall x(A(x) \Rightarrow \exists y B(x, y)). \quad (1)$$

This computation must be done by given subroutines. All input and intermediate values can be kept indefinitely. Quantitative restrictions are not sufficient. It is known that the constructive logic of this type of programming is the usual intuitionistic one; only the formalization of it must be revised to give us better design and better correspondence to both human reasoning and computer programs. The main philosophical principle of the revision is 'contrary to existing prejudices there is no reason to think that what is convenient for a computer is inconvenient for a human and vice versa.'

It was shown in NEPEIVODA (1978a, b, 1979a, b) that *natural deduction* is the most convenient of the existing notions of proof to use in this area. We will present our proofs in well-structured form.

Each auxiliary proof begins by a string

$$]A_1, \dots, A_k, \text{ arbitrary } x_1, \dots, x_l, !x_1, \dots, !x_l$$

(read as 'Let A_1 and ... and A_k for arbitrary x_1, \dots, x_l ; !t read as: 't stops' or 't is meaningful'. The auxiliary proof is marked by a vertical line below the sign $]$ at the beginning of the proof; x_1, \dots, x_l are the bound variables of the auxiliary proof.

If c is to be an auxiliary constant, c is not in our vocabulary and cannot be used outside the auxiliary proof in which it is introduced. An auxiliary proof can be used as a premiss of a rule only once.

Proof rules are generalized in such a way as to exclude superfluous cuts. *Functional formulae* are the formulae of the form ($m > 0, n, k, l \geq 0$)

$$\forall x_1 \dots x_n (A_1 \& \dots \& A_k \Rightarrow \exists y_1 \dots y_l (B_1 \& \dots \& B_m)) \quad (2)$$

shortened to

$$\forall \bar{x} (\bar{A} \Rightarrow \exists \bar{y} \bar{B}). \quad (2')$$

The *functional system* of natural deduction consists of three rules:

(a) The rule of *procedure declaration* (PD):

$$\frac{\begin{array}{c} | A_1, \dots, A_k, \text{ arbitrary } x_1, \dots, x_n, !x_1, \dots, !x_n \\ | \dots \\ | !t_1 \dots !t_l \\ | B_1(y_1 \dots y_l | t_1 \dots t_l) \\ | \dots \\ | B_n(y_1 \dots y_l | t_1 \dots t_l) \end{array}}{\forall \bar{x} (\bar{A} \Rightarrow \exists \bar{y} \bar{B})}$$

(b) The rule of *procedure call* (PC):

$$\frac{\begin{array}{c} | !t_1 \dots !t_n \\ | A_1(x_1 \dots x_n | t_1 \dots t_n) \\ | \dots \\ | A_k(x_1 \dots x_n | t_1 \dots t_n) \\ | \forall \bar{x} (\bar{A} \Rightarrow \exists \bar{y} \bar{B}) \\ | \dots \\ | B_1(\bar{x}, \bar{y} | \bar{t}, \bar{c}) \\ | \dots \\ | B_m(\bar{x}, \bar{y} | \bar{t}, \bar{c}) \\ | !c_1 \dots !c_l \end{array}}{B_1(\bar{x}, \bar{y} | \bar{t}, \bar{c})}$$

Where \bar{c} is the list of *auxiliary constants*.

(c) The rule of *case analysis* (CA)

$$\frac{\begin{array}{c} | \bar{A}_1, \text{ arbitrary } \bar{x}_1, !\bar{x}_1 | \bar{A}_1, \text{ arbitrary } \bar{x}_l, !\bar{x}_l | \bar{A}_{l+1}, \text{ arbitrary } \bar{x}_{l+1}, !\bar{x}_{l+1} \\ | \dots \quad \dots \quad \dots \\ | !\bar{t}_{11}, !\bar{t}_{12} \quad \dots \quad | !\bar{t}_{l1}, !\bar{t}_{l2} \quad \dots \\ | \bar{B}_{\varphi(1)}(\bar{t}_{11}, \bar{t}_{12}) \quad | \bar{B}_{\varphi(l)}(\bar{t}_{l1}, \bar{t}_{l2}) \quad | \text{A} \end{array}}{\exists \bar{y}_1 B_1(\bar{c}, \bar{y}_1) \vee \dots \vee \exists \bar{y}_k B_k(\bar{c}, \bar{y}_k)}$$

where $1 \leq \varphi(i) \leq k$ if $1 \leq i \leq l$, $l \leq n$, $n > 0$.

The conjunctions, disjunctions and lists of variables are considered to be unordered.

PROPOSITION 1. FS is equivalent to the usual intuitionistic logic with partially defined terms.

PROPOSITION 2. The normalization theorem holds for FS.

Example 1. The formal theory of the simple system of programs.

- A1. $\forall x(A(x) \Rightarrow \exists y B(x, y)) \quad \{\chi\}$
- A2. $\forall x(A1(x) \Rightarrow K1(x) \vee K2(x) \vee K3(x) \vee K4(x))$
- A3. $\forall x(A(x) \Rightarrow K1(x) \Rightarrow D(x, \varphi(x)) \& !\varphi(x))$
- A4. $\forall x, y(A2(x) \& K2(x) \& B(x, y) \Rightarrow E(x, \psi(x, y)) \& !\psi(x, y))$
- A5. $\forall x, y(A2(x) \& K3(x) \Rightarrow A(x))$
- A6. $\forall x(K4(x) \Rightarrow \exists y E(x, y)) \quad \{\varkappa\}$
- A7. $\forall x, y(E(x, y) \Rightarrow D(\alpha(x), \beta(x, y)) \& !\alpha(x))$
- A8. $\forall x(A(x) \Rightarrow !\beta(x, y))$

Greek characters after axioms are the names of the functions realizing these axioms.

AIM: $\forall x(A(x) \& A1(x) \& A2(x) \Rightarrow \exists y, z D(y, z))$

PROOF:

| | |
|--|-------------------|
| $\boxed{A(x), A1(x), A2(x)}$ | arbitrary $x, !x$ |
| $B(x, c_1), !c_1$ | (by A1) |
| $K1(x) \vee K2(x) \vee K3(x) \vee K4(x)$ | (by A2) |
| $\boxed{K1(x)}$ | |
| $D(x, \varphi(x))$ | (by A3) |
| $!\varphi(x)$ | |
| $\boxed{K2(x)}$ | |
| $E(x, \psi(x, c_1))$ | (by A4) |
| $!\psi(x, c_1)$ | |
| $\boxed{K3(x)}$ | |
| λ | (by A5) |
| $\boxed{K4(x)}$ | |
| $E(x, c_2)$ | |
| $!c_2$ | (by A6) |
| $D(x, c_3) \vee E(x, c_3)$ | |
| $!c_3$ | |

| | |
|--|---------|
| $JD(x, c_3)$ | |
| $JE(x, c_3)$ | |
| $ D(\alpha(x), \beta(x, c_3))$ | (by A7) |
| $!\alpha(x)$ | |
| $!\beta(x, c_3)$ | (by A8) |
| $D(c_4, c_5), !c_4, !c_5$ | |
| $\forall x(A(x) \& A1(x) \& A2(x) \Rightarrow \exists y, z D(y, z))$ | |

We give some notes about this proof; we can see that we can describe initial functions both in explicit (as α, β) or implicit (α) form. To compare these two formalizations, and for many other purposes, we need the *logical algorithmic language* (LAL), and the corresponding program logic.

There is a very useful logical rule (*Kanger's rule*)

$$\frac{A_1(\bar{t}) \dots A_n(\bar{t})}{A_1(\bar{c}) \dots A_n(\bar{c})}$$

Kanger's rule is admissible for FS. We can simulate it by means of CA with $n = s, l = 0$.

4. A brief sketch of LAL and of intuitionistic program logic

We will give some examples of programs in LAL (first described in NEPEIVODA, 1978b).

Example 2. The program extracted from the proof of Example 1. (Let predicates K1, K2, K3 be computable, and K4 be non-computable.)

```
proc gamma = (ob x) ob y, z:
  (union(ob d, e) c3 = if K1(x) → d: fi(x) □
   K2(x) → e: psi(x, hi(x)) □ K3(x) → error
   else e: kappa(x) fi;
  ob c4, c5 = case c3 in d: x, c3 □ e: alpha(x), beta(x, c3) esac;
  new y: c4, z: c5);
```

LAL is an ALGOL-68-like language. We preserve the block structure of the ALGOL-68 programs, but widen the notions of 'yielding values' and 'declarations'. We discard assignments as they are superfluous. Assignments are, in some sense, like to 'go to'.

We found it striking that logical reasoning yields operator languages rather than functional ones, but it is natural, because operators of

algorithmic languages correspond to cuts. Natural ways of reasoning are roundabout ones.

We can construct an algorithmic logic AFS for proving LAL programs following CONSTABLE (1977). The rules of this logic are the usual ones, the logic is described in NEPEIVODA (1979d).

Example 3. Multiplication by means of addition.

```
proc mult = (nat x, y) nat:
  for i, nat z from 0, 0 to y do i: i+1, z: z+x giving z rof;
```

Example 4. Euclid's algorithm.

```
proc euclid = (nat x, y) nat: for nat u, v from x, y while u ≠ v
  do if u>v then v, u-v else u, v-u fi giving u rof;
```

Example 5. Bubble sorting.

```
proc bubble = ([ ]real a) [ ]real:
  for nat 1, [ ] real b from lwb a, a while i<upb a do
    if b[i]≤b[i+1] then i: i+1, b: b
    else i: i-1, b: b(i := b[i+1], i+1 := b[i]) fi giving b rof.
```

The representation of these known programs in LAL gives us some familiarity with the main features of LAL. Many constructions in LAL are inspired by DIJKSTRA, 1977.

AFS gives us a system for the precise formulation and proof of some theorems which allow us to compare different formalizations and to justify our algorithms for translation. Here we give a list of the main results proved using AFS.

PROPOSITION 3. AFS is a conservative extension of FS.

Let us refer to the notion of 'normal' formula introduced in NEPEIVODA (1978a).

DEFINITION 1. *Normal formulae.*

- (a) Elementary formulae are normal.
- (b) If A, B are normal formulae, then $A \& B$, $\forall x A$ are normal.
- (c) If A_1, \dots, A_k are computable and A_{k+1} is normal, then $A_1 \vee A_k \vee A_{k+1}$ is normal.
- (d) If B is normal, then $A \Rightarrow B$ is normal.
- (e) $\neg A$ is normal.

PROPOSITION 4. *If all the axioms of T are normal, then we have an algorithm for extracting programs from proofs in T such that*

- (a) *The number of steps $t \lfloor P \rfloor$ is less than p^2 , where p is the length of P.*
- (b) *The length of $t \lfloor P \rfloor$ is less than p.*

5. Fuzzy realizability

Let L denote a partially-ordered set. Y is a L -fuzzy set iff $Y \subseteq X \times L$ and $\langle x, \alpha \rangle \in Y$, $\alpha \leq \beta$ implies $\langle x, \beta \rangle \in Y$. $x \in_x Y$ means $\langle x, \alpha \rangle \in Y$.

DEFINITION 2. *Data types.* These are constructed from initial data types **ob1**, **ob2**, ... and **void** by means of operations $m1 \times \dots \times mn$, $m1 \oplus \dots \oplus mn$, $proc(m1, \dots, mn)m$.

DEFINITION 3. A *tower of functionals* is a collection of (fuzzy) functions such that the following conditions hold:

- (a) U is such a function from the set of data types into the class of non-empty L -fuzzy sets that $U(\text{void}) = \{0\}$,

$$U(m1 \times \dots \times mn) = U(m1) \times \dots \times U(mn),$$

$$U(m1 \oplus \dots \oplus mn) = U(m1) \oplus \dots \oplus U(mn).$$

(b) $pr_{m1 \times \dots \times mn}^i$ is a function from $U(m1 \times \dots \times mn)$ into $U(m_i)$ such that $pr_{m1 \times \dots \times mn}^i(\langle x_1, \dots, x_n \rangle) = x_i$.

(c) $inj_{m1 \oplus \dots \oplus mn}^i$ is a standard injection of $U(m_i)$ into $U(m1 \oplus \dots \oplus mn)$.

(d) $join_{m1 \times \dots \times mn}(x_1, \dots, x_n) = \langle x_1, \dots, x_n \rangle$.

(e) $Ev_{proc(m1, \dots, mn)m}$ is a fuzzy partial (multiple-valued) function from $U(proc(m1, \dots, mn)m) \times U(m1) \times \dots \times U(mn)$ into $U(m)$ (evaluation of a function).

(f) $case_{m1 \oplus \dots \oplus mn}^m$ is a fuzzy partial (multiple-valued) function from $U(m1 \oplus \dots \oplus mn) \times U(proc(m1)m) \times \dots \times U(proc(mn)m)$ into $U(m)$ such that $case_{m1 \oplus \dots \oplus mn}^m(x, f1, \dots, fn) = Ev(f_i, x)$ if $x \in U(m_i)$.

(g) $sub_{proc(m1, \dots, mn)m}^{n1, \dots, nk}$ is a function from

$$U(proc(m1, \dots, mn)m) \times$$

$$\times U(proc(n1, \dots, nk)m1) \times \dots \times U(proc(n1, \dots, nk)mn)$$

into $U(proc(n1, \dots, nk)m)$ such that

$$Ev(sub(f, g1, \dots, gn), y1, \dots, yk)$$

$$= Ev(f, Ev(g1, \dots, yk), \dots, Ev(gn, y1, \dots, yk))$$

(h) $S'_{\text{proc}(m1, \dots, mn)m}$ is a function from $U(\text{proc}(m1, \dots, mn)m) \times U(m1) \times \dots \times U(mi)$ into $U(\text{proc}(mi+1, \dots, mn)m)$ such that

$$\text{Ev}(S'(f, x1, \dots, xi), xi+1, \dots, xn) = \text{Ev}(f, x1, \dots, xi, xi+1, \dots, xn).$$

(i) $C_{m1, \dots, mn}$ are such that $\text{Ev}(C, x1, \dots, xn) = x1$.

The notion of tower of functionals is similar to the notion of fully abstract models of typed-calculi except possibly that extensionality may fail.

A notion of morphism of towers is defined naturally. *Foundation* of a tower is a restriction of our functions to a subset of data types.

PROPOSITION 5. *The category of towers given foundation F possesses an initial object.*

This proposition is often used in construction of towers.

Example 6. Suppose we have a system of computers S . Let these computers have different powers. Then some functions can only be computed by some of the computers. On the other hand, different computers can give different results when computing the same function. Let L be the unordered set of all computers from S . Then $\langle x, y \rangle \in_m f$ for all computers m such that f gives y up becoming x .

DEFINITION 4. *Fuzzy realizability.* We define inductively the data type of a formula A and the fuzzy notion ‘ a is the realization of A on the level α ’ (a and $\text{ar}_\alpha A$, resp.).

1. A is normal. $a = \text{void}$ and $\text{Or}_\alpha A \Leftrightarrow A$ is true on the level α .
2. A is $A_1 \vee \dots \vee A_n$, $a = \text{union}(a1, \dots, an)$.
 $\text{cr}_\alpha A \Leftrightarrow c \in_\alpha U(ai) \& \text{cr}_\alpha Ai$.
3. A is $\forall \bar{x} (\bar{B}(\bar{x}) \Rightarrow \exists \bar{y} \bar{C}(\bar{x}, \bar{y}))$.

$$a = \text{proc}(p1, \dots, pn, b1, \dots, bk) r1 \times \dots \times rl \times cl \times \dots \times cm,$$

where $p1, \dots, pn, r1, \dots, rl$ are types of $x_1, \dots, x_n, y_1, \dots, y_l$ resp.

$$\begin{aligned} \text{fr}_\alpha A &\Leftrightarrow \forall x_1, \dots, x_n, b_1, \dots, b_k \forall \beta_1, \dots, \beta_k, \gamma_1, \dots, \gamma_k \in L \\ &\quad (x_1 \in_{\beta_1} U(p1) \& \dots \& x_n \in_{\beta_n} U(pn) \& b1 \in_{\gamma_1} U(b1) \& \dots \& bk \in_{\gamma_k} U(bk)). \\ &\Rightarrow \forall \delta \in L (\delta \leq \alpha \& \delta \leq \beta_1 \& \dots \& \delta \leq \beta_n \& \delta \leq \gamma_1 \& \dots \& \delta \leq \gamma_k \\ &\Rightarrow z(\text{Ev}(f, x_1, \dots, x_n, b1, \dots, bk)) =_\delta z \\ &\quad \& (z)_{l+1} r_\delta C_1((z)_1, \dots, (z)_l) \& \dots \& (z)_{l+m} r_\delta C_m((z)_1, \dots, (z)_m)). \end{aligned}$$

This notion of realizability generalizes many notions similar to Kleene realizability, for example, the theory of constructions (KREISEL, 1965).

PROPOSITION 6. $T \vdash_{\text{FS}} A$ iff A is realizable in each tower where T is realizable.

PROPOSITION 7. Analogous to 6 for AFS.

PROPOSITION 8. If all axioms of T are normal and $E(T)$ is the extraction algorithm for T (as in Proposition 4), then

$$\begin{aligned} T \vdash_{\text{FS}} \forall \bar{x} (\bar{A}(\bar{x}) \Rightarrow \exists \bar{y} B(\bar{x}, \bar{y})) &\text{ iff} \\ T \vdash_{\text{AFS}} \forall \bar{x} (\bar{A}(\bar{x}) \Rightarrow B(\bar{x}, E(p)(\bar{x})) \& !E(p)(\bar{x})) \end{aligned}$$

where p is proof of the first formula in T .

6. Some analogies between programs and proofs

Let us consider more precisely the process of extraction. This process is divided into two consequent parts. First, we mark all active objects in the proof. All output variables of our aim are marked as active immediately. Then we mark all predecessors of those. All auxiliary constants that are used for active variables are marked as active. All disjunctors giving as an active constant or active disjunction are marked as active. All objects that could not be marked as active are passive.

There are direct analogies between the notions of modern algorithmic languages and marked proof.

| | |
|---|--|
| auxiliary active constant | corresponds to simple value |
| " passive | " ghost value |
| active variable | " formal parameter |
| auxiliary proof containing active objects | |
| | corresponds to procedure body or block |
| active disjunction | corresponds to conditions |
| proof | " program |
| active formula | " complex value |
| passive formula | " verification invariant. |

If we introduce induction rules to synthesize loops all these analogies are preserved.

For example, the rule of induction corresponding to primitive recursion is PRI rule:

$$\frac{A(0, t_1, \dots, t_k) \quad | \quad \begin{array}{c} A(n, x_1, \dots, x_k), \text{ arbitrary } n, x_1, \dots, x_k. \\ \dots \\ A(n+1, t_1, \dots, r_k) \end{array}}{A(\varphi, C_1, \dots, C_k)}$$

Here n is a natural number variable, $t_1, \dots, t_k, r_1, \dots, r_k$ are terms. This rule can be translated into

for n , ob1 $x_1, \dots, obk x_k$ from 0, t_1, \dots, t_k to φ
do ... new $n+1, r_1, \dots, r_k$ rof.

In the usual way variables give us the loop parameters and the auxiliary proof gives us the body of the loop.

6. Some logical consequences

Let us consider a tower based on a foundation consisting of natural numbers and primitive-recursive functions. In this tower the following axioms are realizable:

- (a) Quantifier-free properties of primitive-recursive functions.
- (b) Axioms corresponding to the PRI rule.
- (c) The *weak primitive recursive ‘Church’ principle*:

$$\forall x(A(x) \Rightarrow \exists y B(x, y)) \Rightarrow \neg \neg \exists f (\text{PRF}(f) \& \\ \forall x(A(x) \Rightarrow B(x, f(x)))) .$$

PROPOSITION 9. *In the proposed theories we can find a primitive recursive predicate $A(x, y)$ such that*

$\forall x, y(A(x, y) \vee \neg A(x, y)) \& \forall x \neg \neg \exists y A(x, y) \& \neg \forall x \exists y A(x, y)$
holds.

Hint. A expresses a non-primitive-recursive function.

Let us consider an algorithm of constructive deshielding \mathfrak{R} corresponding to fuzzy realizability.

PROPOSITION 10. $T \vdash A$ iff $\mathfrak{R}_{\lfloor T \rfloor} \vdash \mathfrak{R}_{\lfloor A \rfloor}$, where T, A are formulae of the FS language.

PROPOSITION 11. *Given a proof of $\mathfrak{R}_{\lfloor A \rfloor}$ in $\mathfrak{R}_{\lfloor T \rfloor}$ we can reconstruct a proof of A in T in such a way that the length of the proof does not increase.*

PROPOSITION 12. *The program extracted from the proof A in T can be constructed from the one of $\mathfrak{R}_{\lfloor A \rfloor}$ in $\mathfrak{R}_{\lfloor T \rfloor}$ by means of precomputation of repeating expressions.*

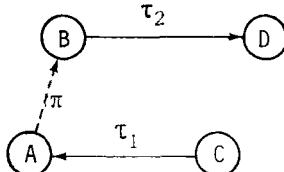
This proposition leads us to the following principle: ‘We must express our functions in an implicit form rather than in an explicit one.’

8. Action planning relevancy logic

Let us consider a fuzzy set of states of world S . *Elementary actions* are fuzzy partial multiple-valued functions from S to S . The set of actions contains an empty action Λ such that $\Lambda(s) = s$ for each s and the set is closed under construction composition.

A first sketch of planning logic is given as follows: We have two variants of each logical connective: relevant and intuitionistic. Relevant connectives are \Rightarrow , *and*, *or*, *not*, intuitionistic ones are expressed as usual. Every formula has its *type of plan*. Relevant connectives can be applied only to formulae of equal type.

We accept the *principle of monotony planning*. If we have a formula $A \Rightarrow B$, then the partial plan corresponding to $A \Rightarrow B$ must use *all* partial plans including the plan of A *precisely once*. For example, if A and B are relevant combinations of elementary formulae, then they denote subsets of S . $A \Rightarrow B$ demands plan Π (i.e. the composition of elementary plans) such that $\Pi(\exists[A]) \subseteq \exists[B]$. $(A \Rightarrow B) \Rightarrow (C \Rightarrow D)$ in the same assumptions demands such a partial plan τ :



Consequently, $(A \Rightarrow B) \Rightarrow (C \Rightarrow D)$ implies

$$(C \Rightarrow A) \ \& \ (B \Rightarrow D).$$

The author wishes to express his gratitude to E. Engeler, P. Martin-Löf, and W. Pratt for their helpful discussions with him.

References

- CONSTABLE, R. L., 1971, *Constructive mathematics and automatic program writers*, IFIP 1971
- CONSTABLE, R. L., 1977, *A constructive programming logic*, IFIP Congr. 77, North-Holland Publ. Co.
- DIJKSTRA, E., 1977, *A systematic programming*
- ERSHOV, A. P., 1977, (Ершов, А. П.) *О сущности процесса трансляции*, Программирование, # 5

- KOGAN, YA. S., 1978, (Коган, Я. С.) *Метод эффективного пополнения доказательств в системе „Логический язык программирования”*, „Искусственный интеллект и автоматизация исследований в математике” (Киев)
- KREISEL, G., 1962, *Foundations of intuitionistic logic*, in: Logic, Methodology and Philosophy of Science, eds. E. Nagel, P. Suppes and A. Tarski (Stanford University Press, Stanford), pp. 192–210
- KREISEL, G., 1975, *Some uses of proof theory for finding computer programs*, Clermont-Ferrant Logic Colloquium
- MANNA, Z., 1977, *Theoretical computer science. An introduction*
- NEPEIVODA, N. N., 1978 a, *A relation between the natural deduction rules and operators of higher level algorithmic languages*, Soviet. Mat. Doklady, vol. 239, #3
- NEPEIVODA, N. N., 1978 b, (Непеивода, Н. Н.) *О построении правильных программ*, Вопросы кибернетики, вып. 46
- NEPEIVODA, N. N., 1979 a, (Непеивода, Н. Н.) *Об одном методе построения правильных программ из правильных подпрограмм*, Программирование, # 1
- NEPEIVODA, N. N., 1979 b, (Непеивода, Н. Н.) *Применение теории доказательств к задаче построения правильных программ*, Кибернетика, # 2
- NEPEIVODA, N. N., 1979 c, *The connections between the proof theory and computer programming*, 6 International Congress of Logic, Hannover, vol. 1, p. 6
- NEPEIVODA, N. N., 1979 d, *A proof. Theoretical comparison of program synthesis and program verification*, 6 International Congress of Logic, Hannover, vol. 1, p. 51

ADMISSIBILITY IN PROOF THEORY; A SURVEY

W. POHLERS

Dedicated to K. SCHÜTTE
on occasion of his 70th birthday

Freiburg—Munich, F.R.G.

I. Why ordinal analysis?

By *ordinal analysis* of a formal system T one usually understands the computation of the order type of the least wellordering which is not provable in T . This ordinal is called the *proof-theoretical ordinal* of T , and we will denote it by $\|T\|$. There has always been criticism against the significance of the ordinal $\|T\|$. Therefore we want to explain in the beginning why we hold ordinal analysis for one of the central tasks of proof theory.

The aim of the “working mathematician” is to detect the true sentences of a given structure M . The way he does it may be described—from a logical point of view—as follows. He axiomatizes M , i.e. he postulates properties which are supposed to be characteristic for M , and then he logically deduces new sentences which are true in M . But normally he will not only speak about objects in M but he also uses functions on M , subsets of M , etc. That means that he erects a set universe V_M above M which is supposed to fulfil certain closure conditions, e.g. comprehension, collection, existence of choice functions, etc. So each of his proofs may be formalized within a formal system producing true sentences of V_M . The proof-theorist is now interested in determining and comparing the strength of those formal systems. But what should he take as a measure for the strength of a formal system T ? We think that the best thing is to test what it can perform with respect to M , i.e. to take the set of sentences in the language of M which are provable in T . To change to a concrete situation we choose M to be the structure N of natural numbers—the simplest infinite structure. N is usually

axiomatized by Peano's axioms. We denote the corresponding formal system by PA (Peano arithmetic).¹ Examples for formal systems involving closure conditions for V_N then are the well-known subsystems of analysis, the systems ID_v, and also subsystems of set theory with N as urelement structure. To measure and to compare the strength of such a formal system T we may proceed in the following way. We define a primitive recursive wellordering $<$ on the natural numbers and determine the least order type α such that $T \vdash \varphi$ iff $\exists \beta < \alpha \text{PA} + \text{TI}(\beta) \vdash \varphi$ for arithmetical φ where $\text{TI}(\beta)$ denotes the scheme of transfinite induction along a segment of $<$ of order type α . If we denote by $\text{TI}(<\alpha)$ all instances of $\text{TI}(\beta)$ with $\beta < \alpha$, this means that T and $\text{PA} + \text{TI}(<\alpha)$ prove the same arithmetical sentences. So α is a measure for the proof-theoretical strength of T in the proposed sense. Obviously, α is also the order type of the least non-provable well-ordering of T if one understands "provable wellordering" in the proper way. So, to us, ordinal analysis seems to be a meaningful problem. The difficulties in solving this problem are twofold. First one has to provide an order relation $<$ in the naturals, which needs the development of a recursive ordinal notation system and second one has to compute $\|T\|$.

It should be mentioned that PA may be replaced by Heyting's arithmetic HA which is based on intuitionistic logic. Clearly, one cannot expect that a classical theory T is conservative over $\text{HA} + \text{TI}(<\|T\|)$ but it turns out that T is conservative over $\text{HA} + \text{TI}(<\|T\|)$ with respect to all arithmetical sentences in the negative fragment. Now it makes sense to ask for the proof-theoretical ordinal of T^i , the formal system T based on intuitionistic logic. In most of the known cases $\|T\|$ and $\|T^i\|$ coincide.² A famous example for this fact are the theories ID_v for v -fold iterated inductive definitions and ID_v^i , the intuitionistic theory for v -fold strictly positive inductive definitions. The proof of $\|\text{ID}_v\| = \|\text{ID}_v^i\|$ (cf. BUCHHOLZ, 1977a; POHLERS, 1977 and 1978) shows that

$$\text{ID}_v \equiv_{-} \text{HA} + \text{TI}(<\|\text{ID}_v\|) = \text{HA} + \text{TI}(<\|\text{ID}_v^i\|) = \text{ID}_v^i,$$

¹ In fact, PA already postulates something for V_N . The scheme of complete induction says that each non-void arithmetic set of V_N has a $<$ -least element. So it would be more natural to choose an axiomatization without induction scheme, for instance, Robinson arithmetic RA. Gentzen's result then shows $\text{PA} \equiv \text{RA} + \text{TI}(< \varepsilon_0)$ (cf. below). All results cited in this paper also hold if PA is replaced by RA.

² This seems to be no longer true for systems like KPN and KPN^i (cf. below). The reason for this may be the lack of good intuitionistic systems for set theory.

i.e. that the classical theory is proof-theoretical reducible to the intuitionistic theory which is constructively meaningful. In the sequel, however, we will neglect all intuitionistic aspects of proof theory. More details about that are in POHLERS (1977) and will appear in FEFERMAN (to appear).

II. Predicative proof theory

The oldest proof-theoretical result in this direction is due to Gentzen who computed $\|\text{PA}\| = \varepsilon_0$. This means that $\text{PA} \equiv \text{PA} + \text{TI}(<\varepsilon_0)$ (cf. footnote 1). His proof was based on a normalization procedure for certain derivations of PA. What prevented the proof from being simple was the fact that PA does not enjoy cut-elimination. The reason for it is the presence of the induction scheme, the only axiom scheme in PA which says something about V_N . It is well known that it is possible to overcome this difficulty by passing to an infinitary system with the so-called ω -rule

$$\vdash \varphi \rightarrow \psi(n) \quad \text{for all } n \in \omega \Rightarrow \vdash \varphi \rightarrow \forall x \psi(x).^3 \quad (\omega)$$

This infinitary system enjoys cut-elimination. But if we add stronger closure conditions for V_N , e.g. comprehension, this property is spoiled again. To make this more visible we will sketch the cut-elimination procedure for the infinitary system. Here, the deductions are ω -branching trees to which an ordinal is canonically connected. We call it the *length* of the tree. By $\left| \frac{\alpha}{n} \varphi \right.$ we will denote that there is a proof tree for φ of length $\leq \alpha$ whose cut formulas are all of complexity $< n$. Our aim is to show that $\left| \frac{\gamma}{n} \varphi \right. \Rightarrow \left| \frac{\delta}{0} \varphi \right.$ where δ is computable from γ . For this it suffices to show that $\left| \frac{\gamma}{n+1} \varphi \right. \Rightarrow \left| \frac{\omega^\gamma}{n} \varphi \right.$, because the rest is a simple iteration. We demonstrate one case:

$$\frac{\dots \left| \frac{\alpha_n}{n} \psi_1 \rightarrow \varphi(m) \wedge \forall x \varphi(x) \dots m \in \omega \right. \left| \frac{\beta_0}{n} \forall x \varphi(x) \vee \varphi(n_0) \rightarrow \psi_2 \right.}{\left| \frac{\alpha}{n} \psi_1 \rightarrow \forall x \varphi(x) \right. \left| \frac{\beta}{n} \forall x \varphi(x) \rightarrow \psi_2 \right.} \left| \frac{\alpha+\beta}{n+1} \psi_1 \rightarrow \psi_2 \right.$$

By cuts we get $\left| \frac{\alpha_{n_0}+\beta}{n+1} \psi_1 \rightarrow \psi_2 \wedge \varphi(n_0) \right.$ and $\left| \frac{\beta_0+\alpha}{n+1} \varphi(n_0) \vee \psi_1 \rightarrow \psi_2 \right.$. By induction hypothesis we get

$$\frac{\left| \frac{\omega^{\alpha_{n_0}+\beta}}{n} \psi_1 \rightarrow \psi_2 \wedge \varphi(n_0) \right. \left| \frac{\omega^\alpha+\beta_0}{n} \varphi(n_0) \vee \psi_1 \rightarrow \psi_2 \right.}{\left| \frac{\omega^\alpha+\beta}{n} \psi_1 \rightarrow \psi_2 \right.}$$

³ We do not know exactly by whom the ω -rule was originally introduced (possibly by Hilbert). The first systematic use of it and its generalizations has been in SCHÜTTE (1960).

since the complexity of $\varphi(n_0)$ is less than n which was the complexity of $\forall x\varphi(x)$ and $\omega^{\alpha_0+\beta}$, $\omega^{\alpha+\beta_0} < \omega^{\alpha+\beta}$. Of course, we have simplified things. A correct induction has to be a bit more subtle (because the bottom most ordinal might be between $\max\{\alpha, \beta\}$ and $\alpha+\beta$) but the example shows how the procedure works in principle. Now if we add a comprehension rule, $\vdash \varphi \rightarrow \psi(S) \Rightarrow \vdash \varphi \rightarrow \exists X\psi(X)$ say, (capital letters are supposed to range over subsets of ω), the above reduction procedure will not longer work since the complexity of $\psi(S)$ may be greater than the complexity of $\exists X\psi(X)$ or $\forall X\psi(X)$. S could be the set $\{x : \exists X\psi(X, x)\}$ for instance. The deeper reason for this is the possible impredicativity of such comprehensions as the example $S = \{x : \exists X\psi(X, x)\}$ shows. One way to overcome the new difficulty is therefore to avoid such impredicative comprehensions. This can be done by defining sets in stages. An arithmetical formula will get the stage O and if $\varphi(x)$ is a formula of stage α then $\{x : \varphi(x)\}$ will be a set of stage α if $\varphi(S)$ with stage $S < \alpha$ is a formula of stage $< \alpha$; the infinite conjunction $\bigwedge \{\varphi(S) : \text{stage } S < \alpha\}$ and disjunction $\bigvee \{\varphi(S) : \text{stage } S < \alpha\}$ are formulas of stage α . (First order operations will not affect the stage of a formula.)

Now we have not only infinite proof trees but also infinitely long formulas which means that their complexity is to be measured by ordinals. The comprehension rule $\vdash \varphi \rightarrow \psi(S) \Rightarrow \vdash \varphi \rightarrow \exists X\psi(X)$ changes into the usual disjunction rule

$$\vdash \varphi \rightarrow \psi(S), \text{ stage } S < \alpha \Rightarrow \vdash \varphi \rightarrow \bigvee \{\psi(S) : \text{stage } S < \alpha\}.$$

But now the complexity of $\psi(S)$ is less than the complexity of

$$\bigvee \{(S) : \text{stage } S < \alpha\}$$

which means that the cut-elimination procedure again works and we get a theorem of the kind

$$\left| \begin{smallmatrix} \alpha \\ \varrho \end{smallmatrix} \right. \varphi \Rightarrow \left| \begin{smallmatrix} f(\alpha, \varrho) \\ 0 \end{smallmatrix} \right. \varphi.$$

At this place we should make clear how an ordinal analysis is obtained from cut-elimination in infinite theories. Suppose we have a formal system T with the property $T \vdash \varphi \Rightarrow \left| \begin{smallmatrix} \alpha \\ \varrho \end{smallmatrix} \right. \varphi^*$ for certain ordinals α, ϱ and an appropriate translation φ^* into the language of the infinitary theory. By cut-elimination it follows $\left| \begin{smallmatrix} \beta \\ 0 \end{smallmatrix} \right. \varphi^*$ for a β depending on α and ϱ . The proof-predicate $\left| \begin{smallmatrix} \beta \\ 0 \end{smallmatrix} \right. \varphi^*$, however, is formalizable in $\text{PA} + \text{TI}(\beta)$. One defines a tree which is locally correct with respect to the inference rules of the infinitary system and obtains its wellfoundedness by an ordinal assignment.

If we have $\vdash_0^\beta \varphi^*\neg$ for an arithmetical formula φ , this derivation has the subformula property, which means that all formulas occurring in it are (translations of) arithmetical ones of bounded complexity, say n . But then we may define a local truth predicate $\text{Tr}_n(\neg\varphi)$ in PA and show $\text{PA} + \text{TI}(\beta) \vdash \vdash_0^\beta \varphi^*\neg \rightarrow \text{Tr}_n(\varphi)$ by induction on β . This gives $T \vdash \varphi \Rightarrow \text{PA} + \text{TI}(\beta) \vdash \varphi$. The least upper bound of the length of all cut-free derivations therefore gives an upper bound for $\|T\|$. To get the opposite direction one shows $T \vdash \text{TI}(\beta)$ for all $\beta < \|T\|$. So the traditional computation of $\|T\|$ gives the desired result even in the sharper sense of I.

It is easy to see that our definition of sets in stages has something to do with the iteration of jumps. More precisely, an iteration of $\omega \cdot \alpha$ jumps will give a set of stage α . If one uses the fact that $(\Delta_1^1\text{-CR})^*$ corresponds to ω^α jumps (due to Feferman) and $(\Delta_1^1\text{-CA})^*$ corresponds to ε_0 jumps (due to Friedman) one can embed those theories into the infinitary system and compute their proof-theoretical ordinals. Schütte and Feferman showed independently that the proof-theoretical ordinal for a theory of autonomously iterated jumps is Γ_0 ($= \theta\Omega_1 0$, cf. Appendix 1). So this ordinal represents the bound for predicativity. It is impossible to exceed it by predicative methods.

III. Admissible proof theory

As we have sketched in Part II, the use of infinitary systems provides a powerful and intelligible tool in predicative proof theory. It therefore has always been our aim to extend this method also to impredicative systems. Since S. Feferman and H. Friedman showed that various impredicative subsystems of analysis are proof-theoretically reducible to theories for iterated inductive definitions, we concentrated on the investigation of those theories.

* $(\Delta_1^1\text{-CR})$ is the following rule

$$\vdash \forall x(\varphi(x) \leftrightarrow \psi(x)) \Rightarrow \vdash \exists X \forall x(\varphi(x) \leftrightarrow x \in X)$$

where φ is a Π_1^1 - and ψ a Σ_1^1 -formula.

$\Delta_1^1\text{-CA}$ is the following scheme

$$\forall x(\varphi(x) \leftrightarrow \psi(x)) \rightarrow \exists X \forall x(\varphi(x) \leftrightarrow x \in X)$$

where φ is a Π_1^1 - and ψ a Σ_1^1 -formula.

The formal system ID_v for v -fold iterated inductive definitions is the system PA augmented by constants P^φ for each X -positive formula $\varphi(X, Y, x, y) \in \mathcal{L}_{\text{PA}}(X, Y)$ and their defining axioms

$$(\text{ID}_v^1) \quad \forall z \prec v \forall x (\varphi(P_z^\varphi, P_{\prec z}^\varphi, x, z) \rightarrow x \in P_z^\varphi),$$

$$(\text{ID}_v^2) \quad \forall z \prec v \{ \forall x [\varphi(\psi, P_{\prec z}^\varphi, x, z) \rightarrow \psi(x)] \rightarrow \forall x (x \in P_z^\varphi \rightarrow \psi(x)) \},$$

where

$$x \in P_z^\varphi : \Leftrightarrow \langle x, z \rangle \in P^\varphi \quad \text{and} \quad x \in P_{\prec z}^\varphi : \Leftrightarrow \exists y \prec z (x \in P_y^\varphi)$$

and \prec is a wellordering on the natural numbers of order type $\geq v$. One often adds the scheme TI(v) to assure that v is a provable wellordering. For $v < \theta\Omega_{\Omega_1}0$ (cf. Appendix 1) this, however, is superfluous.

ID_1 , the non-iterated case, may be regarded as the simplest example of an impredicative system. Ordinal analysis for ID_1 was already obtained by Howard (from above) and Gerber (from below) in 1970 and also later by others using different methods. But there was no obvious generalization to the iterated case.

III. 1. The use of Takeuti's cut-elimination procedure

But fortunately there already existed a cut-elimination procedure for an impredicative formal system. G. TAKEUTI (1967) proved the consistency of $(\Pi_1^1\text{-CA})^6 + \text{BI}$ ⁶ in Gentzen style using a normalization argument for finite derivations. The method was a bit too rough to obtain the proof-theoretical ordinal of $(\Pi_1^1\text{-CA}) + \text{BI}$ but it was possible to refine it and to obtain exact bounds for $(\Pi_1^1\text{-CA})^-$ ⁸. Using this refined procedure and the fact that one iteration of an inductive definition corresponds to one application of $(\Pi_1^1\text{-CA})^-$, we succeeded in computing the proof-theoretical ordinals of ID_v (POHLERS, 1975, 1978; BUCHHOLZ and POHLERS, 1978) and got the conjectured result $||\text{ID}_v|| = \theta\epsilon_{\Omega_v+1}0$ (cf. Appendix 1). But this result was still unsatisfactory for two reasons.

* $(\Pi_1^1\text{-CA})$ is the scheme

$$\exists X \forall x (\varphi(x) \leftrightarrow x \in X) \text{ if } \varphi \text{ is a } \Pi_1^1\text{-formula.}$$

⁷ (BI) is the scheme of classical bar induction. It may be expressed as $\forall X \varphi(X) \rightarrow \varphi(\psi)$ with φ an arithmetical and ψ an arbitrary formula.

⁸ If (S) is an axiom- or inference-scheme we denote its set-parameterfree version by (S)-. In (BI)⁻ no set parameters are allowed in φ but ψ is still an arbitrary formula.

First, it was no extension of the method described in Part II. It is true that in POHLERS (1978) we used an infinitary system with ω -rule to obtain the ordinal analysis but the ω -rule was only needed to deal with the number-theoretical part of $(\Pi_1^1\text{-CA})^-$ and not with $(\Pi_1^1\text{-CA})^-$ itself, contrary to the use of infinitary rules in the analysis of predicative systems. Therefore, the given ordinal analysis did not show what really was going on. The result appeared as a kind of combinatorial miracle and especially it was not possible to get a convincing intuition about the rôle of the higher number classes.⁹

Second, the method stopped too early. We could only treat autonomous iterations of inductive definition leading to the ordinal $\theta\Omega_\alpha 0$, which corresponds to the ordinal $\theta\Omega_1 0$ for autonomously iterated jumps. So there is a parallel feature of iteration of jumps and hyperjumps—an inductive definition corresponds to one hyperjump. But—opposite to predicative theories—the possibilities for iterating hyperjumps are not exhausted by autonomous iterations in impredicative systems. There we are able to define the accessible part of an order relation and may therefore also iterate along accessible parts which leads to essentially stronger systems.

III. 2. Buchholz's $\Omega_{\mu+1}$ -rules

Since inductive definitions and hyperjumps are closely connected, one could try to find an adequate rule for the treatment of inductive definitions by formalizing the hyperjump. This was done by W. Buchholz. He extracted the proof-theoretical content of Howard's paper (HOWARD, 1972) and introduced new rules, called $\Omega_{\mu+1}$ -rules, which formalize the hyperjump. We give a rough sketch of the Ω_1 -rule and an idea how it works in the cut-elimination procedure. Suppose a natural deduction calculus. By $\Pi \vdash \psi$ we denote that Π is a derivation with end-formula ψ , by $\Pi \vdash_0 \psi$ that Π is normal. $D(\psi)$ is the set of all $\Pi \vdash \psi$, $D_0(\psi)$ the set of all $\Pi \vdash_0 \psi$. Recall that P^φ is the constant representing the fixed point of $\varphi(X, x)$. Then the Ω_1 -rule may be posed as follows:

- (Ω_1) If $\Phi: D_0(n \in P^\varphi) \rightarrow D(\psi)$ belongs to a certain class of functions from derivations into derivations, then

$$\langle \Phi(\Pi): \Pi \in D_0(n \in P^\varphi) \rangle \vdash n \in P^\varphi \rightarrow \psi.$$

⁹ Ω is supposed to be the continuous closure of the order-function of the admissible ordinals. Therefore Ω_v is the first ordinal in the $(2+v)$ th constructive number class. Then $\theta\epsilon\Omega_{v+1} 0$ is the collapse of the first ϵ -number greater than Ω_v below $\Omega_1 = \omega_1^{\text{CK}}$.

One easily recognizes the connection to the clause

$$\forall n \in N ([e](n) \in 0 \Rightarrow 3 \cdot 5^e \in 0)^{10}$$

in the definition of Kleene's 0, the hyperjump of N .

The crucial step in the cut-elimination procedure is then the following one

$$\frac{\begin{array}{c} \Pi_0 \quad \dots \Phi(\Pi) \dots \Pi \in D_0(n \in P^\varphi) \\ | \qquad \qquad \qquad \diagdown \qquad \diagup \\ n \in P^\varphi \quad n \in P^\varphi \rightarrow \psi \\ \hline \psi \\ \vdots \end{array}}{,}$$

which may be reduced as follows. We use induction on the length of the derivation. By induction hypothesis it then holds $\tilde{\Pi}_0 \in \widetilde{D}_0(n \in P^\varphi)$. But then $\Phi(\tilde{\Pi}_0) \in D(\psi)$ and again by induction hypothesis $\widetilde{\Phi(\tilde{\Pi}_0)} \in \widetilde{D}_0(\psi)$. Of course, this is only the idea of the rule and the cut-elimination proof belonging to it. The real rule has to take into consideration also open assumptions and sounds therefore more complicated.

Using these new rules, Buchholz reobtained the known results and showed moreover that $||ID_{<_*}|| = \theta_{\epsilon_{\Omega_{\alpha_1+1}}}, 0$, where $ID_{<_*}$ is the system which allows iterations of inductive definitions along the accessible part of an arithmetical defined relation $<$.

III. 3. The method of local predicativity

In spite of the great elegance of Buchholz's work, we still have not been completely satisfied by it since the introduction of the new rules prevents it from being a straightforward generalization of the predicative methods. Apparently the proof predicate $\frac{n}{\sigma} \varphi$ for derivations with $\Omega_{\mu+1}$ -rules has no immediate formalization in $PA + TI(\alpha)$ since one needs inductive definitions to describe that a tree is locally correct with respect to an $\Omega_{\mu+1}$ -rule. So the result $ID_* \equiv PA + TI(< ||ID_*||)$ cannot be obtained without further considerations. It therefore has been our goal to replace the hyperjump by sufficiently often iterated jumps. This goal was gained by the "method of local predicativity" developed in POHLERS (1976) and POHLERS (to appear). The starting point for this method is the fact that the fixed point I_φ

¹⁰ $[e](n)$ is the value of the e th primitive recursive function on n .

of a monotonic inductive definition $\varphi(X, x)$ is obtainable from below as the union of its stages

$$I_\varphi^\xi = \{x : \varphi(I_\varphi^{<\xi}, x)\} \quad \text{with} \quad I_\varphi^{<\xi} = \bigcup \{I_\varphi^n : n < \xi\}.$$

From definability theory we know that $I_\varphi = I_\varphi^{<\Omega_1}$ where $\Omega_1 = \omega_1^{\text{CK}}$ is the first admissible $>\omega$. This definition is at least locally predicative, i.e. predicative in each step, and it is easily formalizable within a formal system using infinitely long disjunctions and conjunctions by defining $n \in I_\varphi^{<\xi} : \Leftrightarrow \bigvee_{n < \xi} n \in I_\varphi^n$ and $n \in I_\varphi^n : \Leftrightarrow \varphi(I_\varphi^{<n}, n)$. By the result of Feferman-Schütte, such a system will not exceed $\theta\Omega_1 0$ but we have not yet axiomatized the fact that the hierarchy of stages I_φ^ξ collapses at Ω_1 . This can be done by introducing a new (finitary) rule

$$(\text{Cl}_{\Omega_1}) \vdash \psi \rightarrow \varphi(I_\varphi^{<\Omega_1}, n) \Rightarrow \vdash \psi \rightarrow n \in I_\varphi^{<\Omega_1}$$

and exactly this rule allows the embedding of ID_1 and therefore makes the system impredicative. Similarly to the situation of the comprehension rule (cf. Part I), this impredicativity is formally manifested by the fact that the premise of the rule has a greater complexity than its conclusion and this is fatal for the cut-elimination procedure. The crucial case will then look like

$$\frac{\frac{\frac{\alpha_0}{\Omega_1} \psi_1 \rightarrow \varphi(I_\varphi^{<\Omega_1}, n)}{\frac{\alpha}{\Omega_1} \psi_1 \rightarrow n \in I_\varphi^{<\Omega_1}} \dots \frac{\frac{\beta_\xi}{\Omega_1} \varphi(I_\varphi^{<\xi}, n) \rightarrow \psi_2 \dots \xi < \Omega_1}{\frac{\beta}{\Omega_1} n \in I_\varphi^{<\Omega_1} \rightarrow \psi_2}}{\frac{\delta}{\Omega_1 + 1} \psi_1 \rightarrow \psi_2}$$

and the standard reduction step will not apply. But fortunately we have some more information. First we observe that the infinitary system has the

Boundedness property

$$\frac{\alpha}{\varrho} \psi_1 \rightarrow \psi(n \in I_\varphi^{<\eta}) \Rightarrow \frac{\alpha}{\varrho} \psi_1 \rightarrow \psi(n \in I_\varphi^{<\alpha}),$$

which means that in $\leqslant\alpha$ steps one only is able to prove something about stages $\leqslant\alpha$.

The proof is by induction on α and the essential case is the inference

$$(\text{Cl}_{\Omega_1}) \quad \frac{\alpha_0}{\varrho} \psi_1 \rightarrow \varphi(I_\varphi^{<\Omega_1}, n) \Rightarrow \frac{\alpha}{\varrho} \psi_1 \rightarrow n \in I_\varphi^{<\Omega_1}.$$

Here the induction hypothesis gives $\frac{\alpha_0}{\varrho} \psi_1 \rightarrow \varphi(I_\varphi^{<\alpha_0}, n)$ which implies $\frac{\alpha}{\varrho} \psi_1 \rightarrow n \in I_\varphi^{<\alpha}$ by an usual \bigvee -introduction inference since $\alpha_0 < \alpha$.

So, if we succeed to get $\alpha_0 < \Omega_1$, we may reduce the crucial cut as follows:

$$\frac{\frac{\alpha_0}{\Omega_1} \psi_1 \rightarrow \varphi(I_\varphi^{<\alpha_0}, n) \quad \frac{\beta_{\alpha_0}}{\Omega_1} \varphi(I_\varphi^{<\alpha_0}, n) \rightarrow \psi_2}{\frac{\delta'}{\Omega_1} \psi_1 \rightarrow \psi_2}.$$

But we need to keep control over the length of the new derivation. In general, δ will be an ordinal $\geq \Omega_1$ while δ' may be less than Ω_1 since we eliminated the Ω_1 -branching inference. That means that we need a collapsing function $D_0: \Omega_2 \rightarrow \Omega_1$. If there is such a collapsing function, we may define a relation $\alpha \ll \beta: \Leftrightarrow \alpha < \beta \wedge D_0\alpha < D_0\beta$ which coincides with the relation $<$ on ordinals $< \Omega_1$ if $D_0 \upharpoonright \Omega_1 = \text{id}_{\Omega_1}$. Now we label the nodes of the derivation tree by ordinals increasing in the sense of \ll instead of $<$. This means that the assigned ordinals remain increasing even after an application of the function D_0 .¹¹ This assignment, however, is impossible in the case of an inference

$$(\wedge) \quad \frac{\beta_\xi}{\alpha} \varphi(I_\varphi^{<\xi}, n) \rightarrow \psi \text{ for all } \xi < \Omega_1 \Rightarrow \frac{\beta}{\alpha} n \in I_\varphi^{<\Omega_1} \rightarrow \psi$$

since there are not Ω_1 -many different $\beta_\xi \ll \beta$. In this case we will require $\beta_\xi < \beta$ and $\beta_\xi \ll \delta$ whenever $\beta \ll \delta$ and $\xi \ll \delta$ holds. Then we may choose $\delta' = f(\delta)$ for any function f such that $\delta \ll f(\delta)$ — in the correct proof $\delta' = \omega^\delta$ will do (cf. POHLERS, to appear, Part I) — since then we have $\alpha_0 \ll \delta$ and $\beta \ll \delta$ which implies $\beta_{\alpha_0} \ll \delta \ll f(\delta)$ and the ordinal assignment in the reduced derivation remains correct. Now suppose that we have a derivation $\frac{\alpha}{\Omega_1} \psi$, where ψ is a formula without negative occurrences of $I_\varphi^{<\Omega_1}$ in it. Such a formula will be called to be of level o . The cut rank Ω_1 assures that there is no negative occurrence of $I_\varphi^{<\Omega_1}$ at all in the derivation which means that no Ω_1 -branching rule is needed. But then the ordinal assignment keeps correct even when we replace all ordinals by their collapses. So we get the

Collapsing property

$$\frac{\alpha}{\varrho} \psi, \varrho \leq \Omega_1 \text{ and } \psi \text{ of level } o \Rightarrow \frac{D_0\alpha}{\varrho} \psi.$$

If we turn back to our example, we see that if we require that $\psi_1 \rightarrow \psi_2$ is of level o , then also $\psi_1 \rightarrow \varphi(I_\varphi^{<\Omega_1}, n)$ will be of level o . But then we get $\frac{D_0\alpha_0}{\Omega_1} \psi_1 \rightarrow \varphi(I_\varphi^{<\Omega_1}, n)$ by the collapsing property, which shows that our assumption $\alpha_0 < \Omega_1$ may be realized.

¹¹ This technique is due to HOWARD (1972).

This method is generalizable to the iterated case (cf. POHLERS, to appear, Part II). But there is another application of this method. The basic idea has been the local predicativity in the definition of I_φ . But there is another striking example of a locally predicative definition: the constructible hierarchy. This hierarchy is easily formalizable within a system with a ramified language where the ramified variables x^α are supposed to range over elements of L_α . The rules corresponding to $(Cl_{\Omega_{\mu+1}})$ then become

$$(Cl_{\Omega_{\mu+1}}) \vdash \forall x^\alpha \exists y^{\Omega_{\mu+1}} \varphi(x, y) \text{ and } \alpha < \Omega_{\mu+1} \Rightarrow \vdash \exists z^{\Omega_{\mu+1}} \forall x^\alpha \exists y \in z \varphi(x, y)$$

where φ is a formula which is Σ -over $L_{\Omega_{\mu+1}}$ (i.e. which does not contain negative quantifiers $\exists x^\beta$ or positive quantifiers $\forall x^\beta$ with $\beta \geq \Omega_{\mu+1}$). The right ordinal assignment assures that this system also has the collapsing property and an easy induction on β shows the

Boundedness property

$$\left| \frac{\beta}{\alpha} \right| \psi_1 \rightarrow \psi(\exists z^{\Omega_{\mu+1}} \varphi) \Rightarrow \left| \frac{\beta}{\alpha} \right| \psi_1 \rightarrow \psi(\exists z^\beta \varphi).$$

Again the crucial step is an inference $(Cl_{\Omega_{\mu+1}})$

$$\left| \frac{\beta_0}{\alpha} \right| \psi_1 \rightarrow \forall x^\alpha \exists y^{\Omega_{\mu+1}} \varphi \Rightarrow \left| \frac{\beta_0}{\alpha} \right| \psi_1 \rightarrow \exists z^{\Omega_{\mu+1}} \forall x^\alpha \exists y \in z \varphi.$$

The induction hypothesis gives $\left| \frac{\beta_0}{\alpha} \right| \psi_1 \rightarrow \forall x^\alpha \exists y^{\beta_0} \varphi$, i.e.

$$\left| \frac{\beta_0}{\alpha} \right| \psi_1 \rightarrow \forall x^\alpha \exists y \in L_{\beta_0} \varphi$$

from which it follows $\left| \frac{\beta}{\alpha} \right| \psi_1 \rightarrow \exists z^\beta \forall x^\alpha \exists y \in z \varphi$ by a usual \exists -introduction inference with witness L_{β_0} .

So the method of local predicativity allows us to eliminate cuts in derivations of (translations of) arithmetical formulas. So we have a very powerful and uniform tool for the ordinal analysis of various systems. One formalizes the fact that T has a model in L to get

$$T \vdash \varphi \Rightarrow \left| \frac{\alpha}{\alpha} \right| \varphi^* \Rightarrow \left| \frac{f(\alpha, \varrho)}{\alpha} \right| \varphi^* \Rightarrow PA + TI(f(\alpha, \varrho)) \vdash \varphi^{12}$$

as sketched in Part II. The details of this sketch are worked out by G. Jager in his thesis (JAGER, 1979). There he also obtained results about subsystems

¹² Since we do not really need a model for all φ but only for arithmetical φ , it may happen that α is less than the smallest β s. t. $L_\beta \vdash T$. Examples are $(\Delta_1^1\text{-CA}) \vdash \varphi \Rightarrow \left| \frac{\alpha}{\alpha} \right| \varphi^*$ with $\alpha < \varepsilon_0$ and also $(\Delta_1^1\text{-CA}) \dashv \varphi \Rightarrow \left| \frac{\alpha}{\alpha} \right| \varphi^*$ with $\alpha < \Omega_{\varepsilon_0}$, while the smallest models are L_{Ω_1} or L_{ε_0} , respectively.

of set theory (cf. Appendix II). A closer look at the scheme ($\text{Cl}_{\Omega_{\mu+1}}$) shows that it is nothing but the axiomatization of the fact that $\Omega_{\mu+1}$ is an admissible ordinal. So we may treat such systems T for which the proof of $L_\alpha \models T$ depends essentially on the admissibility of α . For this reason we called this part of proof theory *admissible proof theory*. In admissible proof theory the close connection between definability (or generalized recursion) theory and proof theory becomes very evident. In proof theory one starts from the knowledge of definability theory and uses the additional information of formal provability to get the desired result. The strongest system which can be handled by this method up to now is $(\Delta_2^1\text{-CA}) + \text{BI}$. One formalizes the fact it has a model at L_{ι_0} , where ι_0 is the first recursively inaccessible ordinal, to get $(\Delta_2^1\text{-CA}) + \text{BI} \vdash \varphi \Rightarrow |_{\iota_0+n}^{\omega+\omega} \varphi^*$ which implies first $|_{\iota_0}^{\omega_n(\omega+\omega)} \varphi^* (\omega_n(\alpha) := \underbrace{\omega \cdot \omega}_n)$ and then $|_0^\alpha \varphi^*$ with $\alpha < \theta \varepsilon_{\iota_0+1} 0$. So $\theta \varepsilon_{\iota_0+1} 0$ is an upper bound for $(\Delta_2^1\text{-CA}) + \text{BI}$ which also can be shown to be exact (cf. JAGER and POHLERS, to appear).

Appendix 1. Ordinal notations

Until now we did not say anything about the other task of a proof-theorist: the development of ordinal notations. Since an even rough sketch of such a development would need too much space we will just say a few words about it to give the reader a better feeling for the ordinals exhibited in the table of Appendix 3. Let \mathcal{A} be the set of admissibles $>\omega$ and $\mathcal{A}_0 = \{0\} \cup \mathcal{A}$. By Ω we denote the continuous closure of the order function of \mathcal{A}_0 . Hence $\Omega_0 = 0$, $\Omega_1 = \omega_1^{\text{CK}}$, $\Omega_\omega = \sup \{\Omega_n : n \in \omega\}$, etc. By $\text{Cl}_\alpha(\beta)$ we denote the closure of $1 \cup \beta$ under $+$, \cdot , Ω and θ_ξ for $\xi < \alpha$, the set of α -critical ordinals is defined by $\text{Cr}(\alpha) = \{\xi : \xi \notin \text{Cl}_\alpha(\xi)\} \cup \mathcal{A}$ and θ_α is supposed to be the order function of $\text{Cr}(\alpha)$. It then follows that for α , $\beta < \theta \Omega_1 0$, $\theta \alpha \beta = \varphi \alpha \beta$ if φ is the usual Bachmann function. If we denote by Λ the first fixed point of Ω which is different from 0, we obtain $\text{Cl}_\Lambda(0) \cap \Omega_1 = \theta \Lambda 0$. This shows that we may denote the segment up to $\theta \Lambda 0$ only using the symbols 0, θ and Ω . As shown by J. Bridge and W. Buchholz, this notation system is recursive. For the investigation of $(\Delta_2^1\text{-CA}) + \text{BI}$, however, this notation system is too weak. In order to obtain a stronger system we denote by ι the set of admissible fixed points of Ω and by ι the continuous closure of its order function. Hence ι_0 is the first recursively inac-

cessible ordinal. Now we define the set $\text{Cl}_\alpha^1(\beta)$ to be the closure of $\Omega_1 \cup \beta$ under $+$, ι and θ_ξ^1 for $\xi < \alpha$, $\text{Cr}^1(\alpha) = \{\xi \in \mathcal{A}: \xi \notin \text{Cl}_\alpha^1(\xi)\} \cup I$ and θ_α^1 as the order function of $\text{Cr}^1(\alpha)$. Then $\text{Cl}_\alpha^0(\beta)$ is the closure of $1 \cup \beta$ under $+$, θ^1 and θ_ξ^0 for $\xi < \alpha$, $\text{Cr}^0(\alpha) = \{\xi: \xi \notin \text{Cl}_\alpha^0(\xi)\} \cup \mathcal{A}$ and θ_α^0 the order function of $\text{Cr}^0(\alpha)$. Then $\theta^1 0\alpha = \Omega_{1+\alpha}$, $A = \theta^1 10$ and for $\alpha, \beta < \theta^0 A_0$, we have $\theta\alpha\beta = \theta^0\alpha\beta$. If we denote by A_1 the first fixed point of ι we obtain $\text{Cl}_{A_1}^0(0) = \theta^0 A_1 0$.

Appendix 2: The spectrum of a formal system

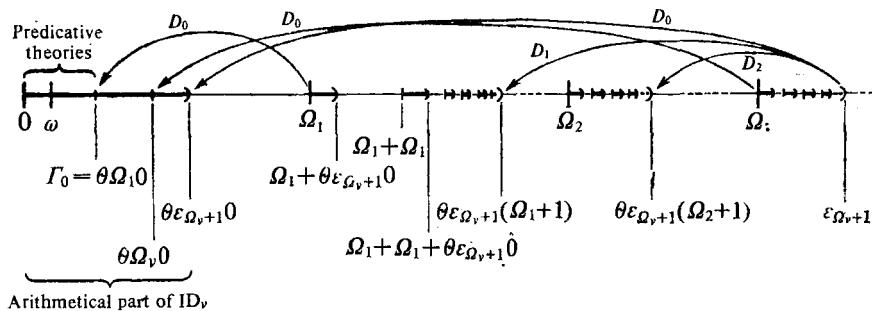
For the elements of the fixed point of an inductive definition there is a canonical norm defined by $|n|_\phi = \min\{\xi: n \in I_\phi^\xi\}$. Then

$$\|\text{ID}_v\| = \sup\{|n|_{\phi_0}: \text{ID}_v \vdash n \in P_0^\phi\}$$

if we denote by ϕ_0 the formula $\phi(X, \emptyset, x, 0)$. It is an obvious question to ask if also the ordinals $\{|n|_{\phi_\mu}: \text{ID}_v \vdash n \in P_\mu^\phi\}$ for $\mu > 0$ are meaningful. Here ϕ_μ means the formula $\phi(X, \tilde{P}_{<\mu}^\phi, x, \mu)$ where $\tilde{P}_{<\mu}^\phi$ is the standard interpretation of $P_{<\mu}^\phi$. If we denote by ID_v^+ the theory ID_v augmented by the scheme $\forall x(\phi(P_v^\phi, x) \rightarrow x \in P_v^\phi)$ we get $\|\text{ID}_v^+\| = \|\text{ID}_v\|$. Therefore we define the spectrum of ID_v to be the set $\text{sp}(\text{ID}_v) := \{|n|_{\phi_\mu}: \mu \leq v \wedge \text{ID}_v^+ \vdash n \in P_\mu^\phi\}$. Then one gets the surprising result

$$\text{sp}(\text{ID}_v) = \text{Cl}_{\varepsilon_{\Omega_v+1}}(0) \cap \varepsilon_{\Omega_v+1}$$

(cf. POHLERS, 1977). Graphically this looks like¹³



¹³ D_0 is the collapsing function mentioned on p. 129. It is defined on $\text{Cl}_A(0)$ and takes values $< \Omega_1$; similarly $D_1: \text{Cl}_A(0) \rightarrow \Omega_2$. Generally one gets collapsing functions $D_\mu: \text{Cl}_A(0) \rightarrow \Omega_{\mu+1}$ (cf. POHLERS, 1978).

One especially gets $\|ID_v\| = \text{sp}(ID_v) \cap \Omega_1$, i.e. $\|ID_v\|$ is the greatest segment contained in $\text{sp}(ID_v)$. Above Ω_1 there are a lot of gaps. It is also possible to define the spectrum for arbitrary formal systems. In all investigated examples it turned out that there is an ordinal τ_T such that

$$\text{sp}(T) = \text{Cl}_{\tau_T}(0) \cap \tau_T$$

and

$$\|T\| = \text{Cl}_{\tau_T}(0) \cap \Omega_1 \quad (= \theta\tau_T 0 \text{ if } \tau_T \geq \Omega_1).$$

This shows that $\|T\|$ only depends on the highest part of the spectrum ¹⁴ and that the arithmetical part of T is quite well characterized by $\text{Sp}(T) \cap \Omega_1$. It is not yet clear if the higher parts of the spectrum have similar properties.

Appendix 3: A table of proof theoretical ordinals

1. Predicative theories

| | System T | $\ T\ $ | τ_T |
|--|--|-------------------------|-------------------------|
| (a) Subsystems of 2nd order arithmetic | RA | ω | ω |
| | PA | ε_0 | ε_0 |
| | $\Pi^0_\infty - \text{CA}$ | $\theta 1\varepsilon_0$ | $\theta 1\varepsilon_0$ |
| | $(\Sigma^1_1 - \text{AC})^r, (\Delta^1_1 - \text{CA})^r$ | ε_0 | ε_0 |
| | $(\Delta^1_1 - \text{CR})$ | $\theta\omega 0$ | $\theta\omega 0$ |
| | $(\Sigma^1_1 - \text{AC}), (\Delta^1_1 - \text{CA})$ | $\theta\varepsilon_0 0$ | $\theta\varepsilon_0 0$ |
| | Autonomously iterated jumps AI | $\theta\Omega_1 0$ | Ω_1 |
| (b) Subsystems of set theory | $(\text{KPN}^+)^r$ | ε_0 | ε_0 |
| | $W\text{-KPN}^+$ | $\theta\varepsilon_0 0$ | $\theta\varepsilon_0 0$ |

¹⁴ Of course, it is possible to construct artificial system for which this becomes false.

2. Impredicative theories

| | System T | $\ T\ $ | τ_T |
|------------------------------|---|--|-------------------------------------|
| (a) First order systems | ID _v | $\theta\varepsilon_{\Omega_{v+1}}0$ | $\varepsilon_{\Omega_{v+1}}$ |
| | ID _{<v, v} limit | $\theta\Omega_v0$ | Ω_v |
| | W-ID _v , v limit | $\theta(\Omega_v\varepsilon_0)0$ | $\Omega_v \cdot \varepsilon_0$ |
| | ID _{<*} | $\theta\varepsilon_{\Omega_{\Omega_1}+1}0$ | $\varepsilon_{\Omega_{\Omega_1}+1}$ |
| (b) 2nd order systems | (Π_∞^0 -CA)+(BI) ⁻ | $\theta(\theta 1\varepsilon_{\Omega_1}+1)0$ | $\theta 1\varepsilon_{\Omega_1}+1$ |
| | (Π_1^1 -CA) ^r | $\theta\Omega_\omega 0$ | Ω_ω |
| | (Π_1^1 -CA) | $\theta(\Omega_\omega \cdot \varepsilon_0)0$ | $\Omega_\omega \cdot \varepsilon_0$ |
| | (Π_1^1 -CA)+BI | $\theta\varepsilon_{\Omega_\omega+1}0$ | $\varepsilon_{\Omega_\omega+1}$ |
| | (Σ_2^1 -AC) ^r , (Δ_2^1 -CA) ^r | $\theta\Omega_\omega 0$ | Ω_ω |
| | (Δ_2^1 -CR)+BI | $\theta\Omega_\omega\omega 0$ | $\Omega_\omega\omega$ |
| | (Σ_2^1 -AC), (Δ_2^1 -CA) | $\theta\Omega_{\varepsilon_0}0$ | Ω_{ε_0} |
| | (Σ_2^1 -AC)+BI, (Δ_2^1 -CA)+BI | $\theta\varepsilon_{\varepsilon_0+1}0$ | $\varepsilon_{\varepsilon_0+1}$ |
| | Autonomously iterated AH hyperjumps | $\theta\Omega_{\Omega_1}0$ | Ω_{Ω_1} |
| (c) Subsystems of set theory | KPN ⁺ | $\theta\varepsilon_{\Omega_1+1}0$ | ε_{Ω_1+1} |
| | KPN ^r | $\theta\varepsilon_{\Omega_v+1}0$ | ε_{Ω_v+1} |
| | KPN ^{<v} | $\theta\Omega_v0$ | Ω_v |

Notes:

1. Superscript ⁻ means that there are no set parameters in scheme (cf. Footnote 8).
2. Superscript ^r in connection with second order systems means that the scheme of complete induction is replaced by the axiom.
3. KPN⁺ is the system described in BARWISE (1975) with urelement structure N (i. e. the PEANO axioms). (KPN⁺)^r means that both the scheme of complete induction and the scheme of foundation are replaced by single axioms. W-KPN⁺ means that the scheme of foundation is replaced by the axiom but the scheme of complete induction is kept. KPN^r is KPN+there exist $< v$ admissibles, KPN^{<v} is the union of all KPN^t, $\xi < v$.

P. S. We have not yet been able to assimilate J. Y. Girard's work in this survey. A comparison of his method with the methods reported here should be very interesting. The references are:

J. Y. GIRARD: Π_2^1 -logic, The Annals of Mathematical Logic, to appear

J. Y. GIRARD: A survey of Π_2^1 -logic (this volume, pp. 89-107)

References

- BARWISE J., 1975, *Admissible sets and structures* (Springer)
- BUCHHOLZ, W., 1974, *Rekursive Bezeichnungssysteme für Ordinalzahlen auf der Grundlage der Feferman-Aczélschen Normalfunktionen θ_α* , Dissertation (Munich)
- BUCHHOLZ, W., 1975, *Normalfunktionen und konstruktive Systeme von Ordinalzahlen*, Springer Lecture Notes in Mathematics, 500
- BUCHHOLZ, W., 1976, *Über Teilsysteme von $\bar{\theta}(\{g\})$* , Archiv für mathematische Logik und Grundlagenforschung, vol. 18
- BUCHHOLZ, W., 1977a, *Some proof theoretical results on the theories ID_v^c , $ID_v^{i(\text{mon})}$, ID_v* , Preprint (Munich)
- BUCHHOLZ, W., 1976, *Eine Erweiterung der Schnitteliminationsmethoden*, Habilitations-schrift (Munich)
- BUCHHOLZ, W., and W. POHLERS, 1978, *Provable wellorderings of formal theories for transfinite iterated inductive definition*, The Journal of Symbolic Logic, vol. 43
- BRIDGE, J., 1975, *A simplification of the Bachmann method for generating large cardinals*, The Journal of Symbolic Logic, vol. 40
- FEFERMAN, S., 1970, *Formal theories for transfinite iterations of generalized inductive definitions and some subsystems of analysis*, Intuitionism and Proof Theory (North-Holland)
- FEFERMAN, S., (ed.) *Iterated inductive definitions and subsystems of analysis: recent proof theoretical studies*, Springer Lecture Notes in Mathematics, to appear
- FRIEDMAN, H., 1970, *Iterated inductive definitions and Σ_2^1 -AC*, Intuitionism and Proof Theory (North-Holland)
- GERBER, H., 1970, *Brouwer's bar theorem and a system of ordinal notations*, Intuitionism and Proof Theory (North-Holland)
- HOWARD, W., 1970, *Assignment of ordinals to terms for type 0 bar recursive functionals* (abstract), The Journal of Symbolic Logic, vol. 35
- HOWARD W., 1972, *A system of abstract constructive ordinals*, The Journal of Symbolic Logic, vol. 37
- JÄGER G., *Beweistheorie von KPN*, Archiv für mathematische Logik und Grundlagenforschung, to appear
- JÄGER G., 1979, *Die konstruktible Hierarchie als Hilfsmittel zur beweistheoretischen Untersuchung von Teilsystemen der Mengenlehre und Analysis*, Dissertation (Munich)
- JÄGER, G., and W. POHLERS, *Admissible proof theory*, to appear.
- KREISEL, G., 1971, *A survey of proof theory II*, 2nd Scandinavian Logic Symposium (North Holland)
- KREISEL, G., 1963, *Generalized inductive definitions*, Stanford Report, Mimeographed (Stanford)
- KREISEL, G., 1968, *A survey of proof theory I*, The Journal of Symbolic Logic vol. 33.
- MARTIN-LÖF, P., 1971, *Hauptsatz for the intuitionistic theory of iterated inductive definitions*, 2-nd Scandinavian Logic Symposium (North-Holland)
- MOSCHOVAKIS, Y. N., 1974, *Elementary induction on abstract structures* (North-Holland)
- PFEIFFER, H. 1970, *Ein Bezeichnungssystem für Ordinalzahlen*, Archiv für mathematische Logik und Grundlagenforschung, vol. 13
- POHLERS, W., 1973, *Eine Grenze für die Herleitbarkeit der transfiniten Induktion in einem schwachen Π_1^1 -Fragment der klassischen Analysis*, Dissertation (Munich)

- POHLERS, W., 1975, *An upper bound for the provability of transfinite induction in systems with N-times iterated inductive definitions*, Springer Lecture Notes in Mathematics 500
- POHLERS, W., 1976, *Eine kanonische Interpretation von ID₁^c*, Lecture at Munster, Mimeo-graphed
- POHLERS, W., 1977, *Beweistheorie der iterierten induktiven Definitionen*, Habilitations-schrift (Munich)
- POHLERS, W., 1978, *Ordinals connected with formal theories for transfinitely iterated inductive definitions*, The Journal of Symbolic Logic, vol. 43
- POHLERS, W., *Cut-elimination for impredicative infinitary systems. Part I: Ordinal analysis for ID₁. Part II: Ordinal analysis for iterated inductive definitions*, Archiv für mathematische Logik und Grundlagenforschung, to appear
- SCHÜTTE, K., 1960, *Beweistheorie* (Springer)
- SCHÜTTE, K., 1977, *Proof Theory* (Springer)
- SIEG, W., 1977, *Trees in Metamathematics*, Thesis (Stanford).
- TAIT, W. W., *Applications of the cut-elimination theorem to some subsystems of classical analysis*
- TAKEUTI, G., 1967, *Consistency proofs of subsystems of classical analysis*, Annals of Mathematics, vol. 86
- TAKEUTI, G., 1975, *Proof theory* (North-Holland)
- TAKEUTI, G., and M. YASUGI, *The ordinals of the systems of second order arithmetic with the provably Δ_2^1 -comprehension rule and with the Δ_2^1 -comprehension axiom respectively*, Japanese Journal of Mathematics, vol. 41
- ZUCKER, J., 1973, *Iterated inductive definitions, trees and ordinals*, Springer Lecture Notes in Mathematics 344

For a more detailed bibliography of predicative proof theory cf. the bibliography given in Kreisel (1971; 1968), Schütte (1960, 1977), and Takeuti (1975).

SYMPOSIUM ON CONSTRUCTIVITY IN MATHEMATICS: INTRODUCTORY REMARKS

R. O. GANDY

Mathematical Institute, Oxford, England

In 1918 H. Weyl bet G. Polya that in 20 years time two typical theorems of classical mathematics (the theorem of the least upper bound and the theorem that every infinite set includes a countable set) would be regarded as meaningless or false; or if, alternatively, they were accepted as true, that would be because the terms involved would have been given an entirely new and unforeseeable interpretation.¹ Weyl could not have been more wrong. Sixty years later not only are those theorems taught to every young mathematician, but also Cantor's conception of a never-ending hierarchy of larger and larger actually infinite sets has been used increasingly in a wide variety of mathematical disciplines.

Mathematics is largely a product of our creative imagination; Cantor wanted pure mathematics to be known as *free* mathematics. And the job of philosophers of mathematics is to understand mathematics, not to change it. Thus the task before us is not to try and settle, by rehashing familiar arguments, a dispute between classical and constructive mathematics. (I recall that Einstein once asked a friend 'What is this frog-and-mouse battle between Hilbert and Brouwer?'.) Rather we should explore the ways in which constructive mathematics—which has also developed considerably since 1918—can contribute to mathematics as a whole. I will briefly describe some of these ways.

It is a platitude that one of the chief advantages of a constructive proof is that it provides more information than does a classical proof of a related result. But this, like other platitudes, seems to crumble away when one

¹ The exact terms of the bet were reproduced in 'The Mathematical Intelligencer', No. 1, 1977, Springer Verlag.

examines it more closely. The best known instances of the calculation of explicit number-theoretic bounds have either used classical methods or the proof-theoretic analysis of classical proofs.² Although E. Bishop insists that the important thing about his work in constructive analysis is that it always yields computable solutions, yet he does not, as far as I know, consider the application of his work to numerical analysis. Indeed I know of few cases where the response to a need for further information has been to search for a constructive proof.³ Here then is an under-exploited application of constructive mathematics.

[Added after the symposium: Both Martin-Löf's contribution and the impassioned interventions by N. N. Nepeivoda stressed the connections between constructive proofs and programmes. Martin-Löf shows how the objects of constructive analysis (as formulated in his theory of types) can be thought of as programmes for a (potentially infinite) computer. Nepeivoda, partly inspired by Kreisel (1977b) in his contribution to this volume, explores ways in which constructive proofs can be used to construct programmes. Most promisingly he pays attention to the connection between restricted methods of proof and degrees of complexity of computation.]

Another interesting line of investigation is described in J. M. E. Hyland's contribution. The classical analysis of, say, a differential equation which models some physical situation often leads to extreme discontinuities between the solutions corresponding to slight variation of the boundary conditions. But in the real world violent discontinuities are, fortunately for us, seldom apparent. The applied mathematician may account for the discrepancy by reference to neglected dissipative and elastic forces or to the need for a statistical treatment. But he may be unable to give any *general* account of the results to be expected if efforts are made to remedy the neglect or to use statistical methods. Solutions obtained by constructive methods will however exhibit some degree of continuity in the parameters. Particularly interesting is Hyland's suggestion that in complex situations one should look out for those mathematical notions which can be handled constructively, because they may prove to have a physical interpretation in terms of what is experimentally observable. (In the simplest situation—measurement of real valued quantities—the connection between construction and observation is well known.)

Finally, I would like to mention applications of constructivist thought to the philosophy of mathematics as a whole. Let it first be said that purely

² KREISEL (1977a) reviews and discusses these calculations.

³ One exception is D. S. Bridges' work on polynomial approximation; see BRIDGES (1979, pp. 160–163) and further forthcoming papers.

negative criticisms may be of value. Freely proceeding mathematicians are often unclear about exactly what they have created and how it is to be understood. Bishop Berkeley was right to point out contradictions in the handling of vanishingly small quantities; he was also right in not doubting the soundness of the *results* obtained by the infinitesimal calculus. (And we now know how to reformulate consistently the theory of infinitesimals.) One way of formulating Brouwer's original critique is to say that classical quantification theory over infinite collections only provides a circular explanation of the way in which its statements are to be understood: ' $\forall n \varphi(n)$ ' means that φ holds for *all* natural numbers; Tarski's definition of truth is totally unilluminating. So I ask the question: is it possible to make use of the attractive and sophisticated theories of meaning which have been developed for constructive logic by DUMMETT (1974, 1977), PRAVITZ (1977), HANCOCK and MARTIN-LÖF (1975) to clarify, to illuminate, the meaning of statements of classical mathematics? Simple interpretations of classical logic in intuitionistic logic do not serve. (My own belief is that the *ultimate* account of ways in which we understand mathematical statements will be given in ultra-finitistic terms.)

References

- BRIDGES, D. S., 1979, *Constructive functional analysis*, Research Notes in Mathematics No. 28 (Pitman, London)
- DUMMETT, M. E., 1974, *The philosophical basis of intuitionistic logic*, in: Logic Colloquium '73, ed. H. E. Rose and J. C. Shepherdson (North-Holland Pub. Co., Amsterdam)
- DUMMETT, M. E., 1977, *Elements of intuitionism* (Oxford)
- HANCOCK, P., and P. MARTIN-LÖF, 1975, *Syntax and semantics of the language of primitive recursive functions*, Preprint No. 3, Dept. of Mathematics, University of Stockholm
- KREISEL, G., 1977a, *On the kind of data needed for a theory of proofs*, in: Logic Colloquium '76, ed. R. O. Gandy and J. M. E. Hyland (North-Holland Pub. Co., Amsterdam)
- KREISEL, G., 1977b, *Some uses of proof theory for finding computer programmes*, in: Colloque International de Logique, Clermont-Ferrand 1975 (Editions du C. R. N. S., Paris)
- PRAVITZ, D., 1977, *Meaning and proofs: on the conflict between classical and intuitionistic logic*, *Theoria*, vol. 43, pp. 2-40

APPLICATIONS OF CONSTRUCTIVITY

J. M. E. HYLAND

(D. P. M. M. S.) *Cambridge, England*

This paper contains a sketch of some proof-theoretical results concerning constructive mathematics and indications how results may be used both to understand results in pure mathematics and (more optimistically) as a guide in discovering physically significant results. The proof-theoretic results are concerned with notions of (local) continuity in parameters. Interest in such questions goes back at least as far as *Hadamard's principle*: in order that differential equations be physically meaningful, they should have solutions uniquely determined by the initial and boundary conditions and should be stable (small changes in the data produce only small changes in solutions). Of course, as has been popularized by Catastrophe Theory, there are many instabilities in the real world; so no crude interpretation of Hadamard's principle is plausible (or useful).

The claims that underly this paper are

- (i) that for all practical purposes questions about continuity in parameters are questions about the constructivity of arguments, and
- (ii) that constructivity may operate as a useful heuristic principle in the application of mathematics.

As regards (i), what we give are results deriving continuity from constructivity, and simple examples of this connection. It is inevitably difficult to *show* that "natural" continuity results will always be obtainable by constructive arguments, though results from HYLAND (1977) can be used to give some plausibility to this. In Section 3, I suggest a test case for claim (i). As regards (ii), the hope is to make use of experience of what are good constructive definitions, in particular in the context of topos theory. Clearly considerable work will be needed to realize this hope.

The idea that there is some connection between constructivity and continuity is an old one (Brouwer's Theorem that all functions from reals

to reals are continuous). But until recently, there was no presentation of useful formal results. HYLAND (1977) contained model-theoretic results giving connections both ways between constructivity and continuity. Independently, BEESON (1977) produced general proof-theoretic results connecting constructivity with his notion of stability. This paper merges Beeson's work with the fundamental distinction between systems with and without choice principles alluded to in HYLAND (1977). No attempt is made here to be comprehensive; the sole aim is to give the flavour of the area.

1. Formal systems

The proof-theoretic results with which we shall be concerned go through particularly smoothly for a basic system of intuitionistic type theory (with extensionality), or equivalently (see for example FOURMAN, 1977) in the context of topos theory (for which see JOHNSTONE, 1977). However, the reader may well prefer something more down to earth, so we consider a system with just two levels of higher types. Specifically our system is based on intuitionistic predicate logic, and has

- (a) *types* closed under pairing, with basic type N for the natural numbers, and with two levels of power types (so e.g. $P(N)$ and $P(N \times P(N))$);
- (b) *term forming operations* of application, pairing and unpairing among the types;
- (c) *axioms*, full second order “Peano” axiomatization for N , extensionality and full comprehension.

(For ease of expression of ordinary mathematics, comprehension may be taken to include the introduction of definable subtypes; indeed, one would naturally make many conservative extensions of this system, allowing for example the direct formation of function spaces $A \rightarrow B$ and power types $P(A)$, restricting A to be of level 0 or 1.) In this system we will have definable types R for the Dedekind reals (as in FOURMAN and HYLAND, forthcoming, or JOHNSTONE, 1977), $R \rightarrow R$ for functions from reals to reals, $Cts(R, R)$ for continuous such functions and so on. If we needed higher levels of types, we could add them.

For the purpose of formalizing mathematical practice, the basic system may be augmented in two distinct ways.

- (1) We may add an axiom stating the compactness of Cantor space 2^N (i.e. the intuitionist's Fan Theorem). From this axiom we readily obtain

the compactness of the unit interval, the uniform continuity, and hence integrability of continuous functions from the unit interval to R , and also general ways of transforming local into global properties. It is important that we do not use the Cauchy reals in this system: they are not complete with respect to the usual uniformity. Monotonic bar induction could be added (though not to much effect), but *axioms of (dependent or countable) choice are excluded*. Let S denote this system, *the system without choice principles*.

(2) We may add to S the axiom DC of dependent choice. Then (see FOURMAN and HYLAND, forthcoming) the Cauchy and Dedekind reals may be identified. Approximation arguments involving choice become available and a theory of Lebesgue integration developed more or less along traditional lines (see BISHOP and CHENG, 1972). Let S^* denote this system, *the system with choice principles*.

The systems S and S^* are subsystems of systems formalizing classical mathematics. They are based on no philosophical analysis, and are simply designed to stay as close as possible to usual mathematical practice. But it seems worth commenting on the relation between S , S^* and other constructive approaches.

(A) Forget his remarks hinting at an intensional interpretation, and Bishop's mathematics (see BISHOP, 1967, or BRIDGES, 1979) can be formalized straightforwardly in S^* . So clearly there are connections with extensional systems designed for this (e.g. that of FRIEDMAN, 1977). Our set-theoretical apparatus has been restricted for simplicity.

(B) If from the system of KLEENE and VESLEY (1965), one drops the axiom of *continuous* choice (§ 7 of Chapter 1) which is at variance with classical mathematics, one obtains a subsystem of $S +$ bar induction. Most of the development in KLEENE and VESLEY (1965) depends only on the fan theorem and numerical choices, and so can be formalized in S^* .

(C) KLEENE and VESLEY (1965) is similar in spirit to our systems in that there is no consideration of different kinds of sequences. There is naturally a much greater difference between our systems and the theory of choice sequences as exposed in KREISEL and TROELSTRA (1970). The difference can best be indicated by observing that the axiom of *bar continuity* may be analyzed as an amalgum of principles of *choice*, *continuity*, *effectivity* (using lawlike sequences) and *bar induction* (in that continuous functionals are in the inductively defined set K of KREISEL and TROELSTRA, 1970). Of this all we have is the weaker Fan Theorem in S and weaker choice principles in S^* .

(D) Without compactness, S^* would be a system consistent with Church's Thesis, and would formalize effective analysis (where all objects are effectively given), but we do not consider that here.

2. Continuous dependence and stability

The results given below extend to the general situation considered in BEESON (1977), namely dependence of values from a separable metric space on parameters from a complete separable metric space, and to other cases. But we restrict to the case of real values and parameters.

DEFINITION. Suppose $\forall x \in R. \exists y \in R. \Phi(x, y)$. We say that *locally* y depends continuously on x (i.e. can be chosen continuously) if and only if $\forall x \in R. \exists$ neighbourhood N_x of $x. \exists f \in \text{Cts}(N_x, R). \forall x' \in N_x. \Phi(x', f(x'))$.

We say (following BEESON, 1977) that y is *stable* in x (i.e. can be chosen stably) if and only if

$$\begin{aligned} \forall x \in R. \exists y \in R (\Phi(x, y) \\ \wedge \forall \text{ neighbourhood } N_y \text{ of } y. \exists \text{ neighbourhood } N_x \text{ of } x. \\ \forall x' \in N_x. \exists y' \in N_y. \Phi(x', y')) \end{aligned}$$

Remarks. 1. Clearly, if globally y depends continuously on x , it does so locally, and if it does so locally, then y is stable in x . If y is uniquely determined by x , then the converse implications hold, but in general they do not (see Examples (1) and (2) below).

2. A solution locally continuous in parameters clearly satisfies the continuity intentions behind Hadamard's principle. It may be too strong, but stability appears to be too weak (see Example (4), though this is not a good physical example).

THEOREM. (i) (HAYASHI; HYLAND) *If $\forall x \in R. \exists y \in R. \Phi(x, y)$ is provable in S , then locally y is continuous in x .*

(ii) (BEESON) *If $\forall x \in R. \exists y \in R. \Phi(x, y)$ is provable in S^* , (and if $\forall x \in R. \{y \mid \Phi(x, y)\}$ is closed is provable in S^*), then y is stable in x .*

A proof of (i) for the basic system (i.e. S without compactness of $2^{\mathbb{N}}$) is in HAYASHI (preprint). I hope to publish my independent (but later) proof based on category-theoretic ideas. (ii) is in BEESON (1977); I doubt whether the condition in brackets is really essential.

3. Applications

To facilitate appreciation of the result above, I first give some

Examples from elementary mathematics.

(1) *Distinction between global and local continuity.* Consider the solution of the cubic equation

$$x^3 - 3x = a,$$

in reals for real parameter a . A diagram convinces one that there is no globally continuous solution, but that there are locally continuous ones. The equation can be proved to have a solution in S (split initially into cases: $a > -1/2 \vee a < 1/2$), so local continuity is a consequence of (i) of our Theorem.

(2) *Distinction between local continuity and stability.* Consider the solution of $z^2 = c$ in complex numbers for complex parameter c . There is no continuous solution of this equation in any neighbourhood of 0 (look at $\arg z$ —there is a homotopy obstruction)! Hence by part (i) of our Theorem, we cannot show the existence of a solution in S . *A fortiori*, the fundamental theorem of algebra is not provable in S . However, a constructive version is provable in S^* (as in BISHOP and CHENG, 1972), so in accordance with part (ii) of the Theorem, z can be chosen stably in c .

Warning: (1) and (2) should not be confused. (1) is frequently quoted to show that arbitrary choice principles cannot be used constructively with extensionality, and for this (2) would do just as well. But there is an important difference. Analogous to (2) is the solution of $x^3 + ax + b = 0$ in reals for real parameters a and b . The solutions form the fold and a solution cannot be chosen continuously in any neighbourhood of $a = b = 0$.

OPEN PROBLEM. One can express in a good constructive way (i.e. by a coherent formula) a condition $\text{Sep}(f)$ on the degree n polynomial f , classically (and indeed in every Grothendieck topos) equivalent to the existence of at least one simple root of f . The schema

$$\text{Sep}(f) \rightarrow \exists z. f(z) = 0,$$

expresses the separable closure of the complex numbers. There is no known proof of this in our basic system or in S . Since the simple root is certainly locally continuous in the coefficients, this seems a good *test case* for the claim that natural continuity results can always be obtained by considerations of constructivity.

(3) Using a simple extension of our main theorem, the difference between S and S^* can be detected in the proof in S^* that every Dedekind real is Cauchy! This gives for each element of R a sequence of rationals (element of $N \rightarrow Q$) converging to it. But there are no non-trivial continuous maps from open sets in R to $N \rightarrow Q$. Here $N \rightarrow Q$ may have the product topology with Q given either the subspace topology induced from R , or the discrete topology (so $N \rightarrow Q$ is isomorphic to Baire space). In either case, we will have stability on the basis of (ii) of the Theorem.

(4) *Computational but non-physical significance of stability.* Cantor's theorem on the uncountability of R formulated constructively states

$$\forall f \in N \rightarrow R. \exists x \in R. \forall n. x \neq f(n),$$

where \neq is the intuitionistic apartness on R . x cannot be chosen (locally) continuously in f (an amusing exercise), so Cantor's theorem cannot be proved in S . But it can easily be proved in S^* , and indeed the stability of x is a triviality. It is hard (for me) to imagine stability of this kind being physically significant.

Next still at the very simplest level, I give the obvious

Application to differential equations. If a differential equation

$$\frac{dy}{dx} = f(x, y)$$

is such that f satisfies a Lipschitz condition on y , then for any a and b , there is an interval I containing a and a (continuous and) differentiable function $g: I \rightarrow R$ with $g(a) = b$ and satisfying the differential equation throughout I .

The standard argument for this may readily be formalized in S . So by a suitable generalization of our Theorem, I and g may be found continuously in f , a Lipschitz constant K , a and b . The only non-global feature is the interval I depending on K , so fix K and we have a quite general continuity of the solution g in the initial condition and the function f determining the equation. (Naturally, continuity here is with respect to the topology of continuous convergence or the compact open topology.)

Finally I give a brief discussion of

Applications to variational problems. This is a large area. There are many classical problems, for example Steiner's problem and Plateau's problem (Beeson has announced a paper devoted to the latter, BEESON, forthcoming).

To get a feel for what goes on, consider the simple result in analysis which provides the conceptual background for variational problems: a continuous function on the unit interval has a maximum (as is provable in S) and attains it (which is essentially non-constructive). By an extension of our Theorem (i), the maximum depends continuously on the function. But it is not possible to choose a point at which the maximum is attained stably in the function (consider $e^{ax}\sin(x)$ in $[0, 4\pi]$ and vary the parameter a through 0); so attainment of bounds is not provable in S^* . (The problem is connected with non-uniqueness, but it has a different effect here than it did above.) Continuity of the extremal value, but instability of the position of attainment, is a persistent feature of variational problems. Naturally, one then turns to the physically significant question of continuity of relative (strict) extrema, where there these exist. Such questions are difficult, and answers vary from problem to problem. At this time, I cannot give a constructive account of all the examples that come to mind.

I conclude this paper by making some very tentative remarks about

Possible heuristic value in science. Attempts to model many physical problems give rise to differential equations whose dynamics depends sensitively on initial conditions: for example every trajectory in some bounded region may be Liapunov unstable. The term *chaos* has been applied to extreme situations of this sort, and they are the subject of much research at this time. Some results are known concerning the topological structure of attracting sets which occur in particular situations, and much more is plausibly indicated on the basis of computer simulation. But the detailed structure of the flow is too complicated for any useful description. This is because of the sensitivity of the dynamics, or as I wish to say, because of the impossibility of a constructive treatment of the flow (in the large).

A recent paper by SHAW (preprint) gives a view of this situation based on a sophisticated interpretation of Hadamard's principle. His idea is that for physical applications the detailed structure does not matter; for example, exactly what the flow pattern is, is not of importance in turbulence, it is changing all the time; but certain general features of it remain the same. Again the exact shape of the record of one's heartbeat on an oscilloscope does not matter, though certain general features may be of considerable significance. So what are important are characteristic properties of a dynamical system which are continuously dependent on parameters; or which we can handle constructively! One such for which Shaw gives an interesting discussion is the rate of creation or loss of information in the flow. There

must be others. Certainly for particular dynamical systems, there would be interest in regions where there were “approximately periodic orbits” of “approximately the same period”, if this qualitative situation is stable. And we could expect such qualitative properties to be established constructively. Of course, a physicist does not need to be told by logicians what properties of a system are of importance. But our physical intuition of stability is fallible, so illustrations and applications mentioned in this paper may be of some use.

References

- BEESON, M. J., 1977, *Principles of continuous choice and continuity of functions in formal systems for constructive mathematics*, Annals of Mathematical Logic, vol. 12, pp. 249–322
- BEESON, M. J., *Plateau's problem and constructive mathematics* (forthcoming)
- BISHOP, E. A., 1967, *Foundations of constructive analysis* (McGraw-Hill)
- BISHOP, E. A., and H. CHENG, 1972, *Constructive measure theory*, Memoir of the American Mathematical Society, No. 116
- BRIDGES, D. S., 1979, *Constructive functional analysis* (Pitman)
- FOURMAN, M. P., 1977, *The logic of topoi*, in: Handbook of Mathematical Logic (North-Holland)
- FOURMAN, M. P., and J. M. E. HYLAND, *Sheaf models for analysis*, Proceedings of the Durham Symposium on Sheaves and Logic (forthcoming)
- FRIEDMAN, H., 1977, *Set theoretic foundations for constructive analysis*, Annals of Mathematics (2), vol. 105, pp. 1–28
- HAYASHI, S., *Derived rules related to the topos theoretic analysis in the intuitionistic higher order arithmetic*, preprint
- HYLAND, J. M. E., 1977, *Aspects of constructivity in mathematics*, in: Logic Colloquium '76 (North-Holland)
- JOHNSTONE, P. T., 1977, *Topos theory* (Academic Press)
- KLEENE, S. C., and R. E. VESLEY, 1965, *The foundations of intuitionistic mathematics* (North-Holland, Amsterdam)
- KREISEL, G., and A. S. TROELSTRA, 1970, *Formal systems for some branches of intuitionistic analysis*, Annals of Mathematical Logic, vol. 1, pp. 229–387
- SHAW, R., *Strange attractors, chaotic behaviour and information flow*, preprint

CONSTRUCTIVE MATHEMATICS AND COMPUTER PROGRAMMING

PER MARTIN-LÖF

University of Stockholm, Stockholm, Sweden

During the period of a bit more than thirty years that has elapsed since the first electronic computers were built, programming languages have developed from various machine codes and assembly languages, now referred to as low level languages, to high level languages, like FORTRAN, ALGOL 60 and 68, LISP and PASCAL. The virtue of a machine code is that a program written in it can be directly read and executed by the machine. Its weakness is that the structure of the code reflects the structure of the machine so closely as to make it unusable for the instruction of any other machine and, what is more serious, very difficult to understand for a human reader, and therefore error prone. With a high level language, it is the other way round. Its weakness is that a program written in it has to be compiled, that is, translated into the code of a particular machine, before it can be executed by it. But one is amply compensated for this by having a language in which the thought of the programmer can be expressed without too much distortion and understood by someone who knows next to nothing about the structure of the hardware, but does know some English and mathematics. The distinction between low and high level programming languages is of course relative to available hardware. It may well be possible to turn what is now regarded as a high level programming language into machine code by the invention of new hardware.

Parallel to the development from low to high level programming languages, there has been a change in one's understanding of the programming activity itself. It used to be looked (down) upon as the rather messy job of instructing this or that physically existing machine, by cunning tricks, to perform computational tasks widely surpassing our own physical powers,

something that might appeal to people with a liking for crossword puzzles or chess problems. But it has grown into the discipline of designing programs for various (numerical as well as nonnumerical) computational tasks, programs that have to be written in a formally precise notation so as to admit of automatic execution. Whether or not machines have been built or compilers have been written by means of which they can be physically implemented is of no importance as long as questions of efficiency are ignored. What matters is merely that it has been laid down precisely how the programs are to be executed or, what amounts to the same, that it has been specified how a machine for the execution of the programs would have to function. This change of programming, which DIJKSTRA (1976, p. 201) has suggested to fix terminologically by switching from computer science to computing science, would not have been possible without the creation of high level languages of a sufficiently clean logical structure. It has made programming an activity akin in rigour and beauty to that of proving mathematical theorems. (This analogy is actually exact in a sense which will become clear below.)

While maturing into a science, programming has developed a conceptual machinery of its own in which, besides the notion of program itself, the notions of data structure and data type occupy central positions. Even in FORTRAN, there were two types of variables, namely integer and floating point variables, the type of a variable being determined by its initial letter. In ALGOL 60, there was added to the two types **integer** and **real** the third type **Boolean**, and the association of the types with the variables was made both more practical and logical by means of type declarations. However, it was only through HOARE (1972) that the notion of type was introduced into programming in a systematic way. In addition to the three types of ALGOL 60, there now appeared types defined by enumeration, Cartesian products, discriminated unions, array types, power types and various recursively defined types. All these new forms of data types were subsequently incorporated into the programming language PASCAL by WIRTH (1971). The left column of the table on the next page, which shows some of the key notions of programming and their mathematical counterparts, uses notation from ALGOL 60 and PASCAL.

As can be seen from this table, or from recent programming texts with their little snippets of set theory prefaced to the corresponding programming language constructions, the whole conceptual apparatus of programming mirrors that of modern mathematics (set theory, that is, not geometry) and yet is supposed to be different from it. How come? The reason for this

| Programming | Mathematics |
|--|--------------------------------------|
| program, procedure, algorithm | function |
| input | argument |
| output, result | value |
| $x := e$ | $x = e$ |
| $S_1; S_2$ | composition of functions |
| if B then S_1 else S_2 | definition by cases |
| while B do S | definition by recursion |
| data structure | element, object |
| data type | set, type |
| value of a data type | element of a set, object of a type |
| $a : A$ | $a \in A$ |
| integer | \mathbb{Z} |
| real | \mathbb{R} |
| Boolean | $\{0, 1\}$ |
| (c_1, \dots, c_n) | $\{c_1, \dots, c_n\}$ |
| array $[I]$ of T | $T^I, I \rightarrow T$ |
| record $s_1:T_1; s_2:T_2$ end | $T_1 \times T_2$ |
| record case $s : (c_1, c_2)$ of $c_1:(s_1:T_1); c_2:(s_2:T_2)$ end | $T_1 + T_2$ |
| set of T | $\{0, 1\}^T, T \rightarrow \{0, 1\}$ |

curious situation is, I think, that the mathematical notions have gradually received an interpretation, the interpretation which we refer to as classical, which makes them unusable for programming. Fortunately, I do not need to enter the philosophical debate as to whether the classical interpretation of the primitive logical and mathematical notions (proposition, truth, set, element, function, etc.) is sufficiently clear, because so much is at least clear, that if a function is defined as a binary relation satisfying the usual existence and unicity conditions, whereby classical reasoning is allowed in the existence proof, or a set of ordered pairs satisfying the corresponding conditions, then a function cannot be the same kind of thing as a computer

program. Similarly, if a set is understood in Zermelo's way as a member of the cumulative hierarchy, then a set cannot be the same kind of thing as a data type.

Now, it is the contention of the intuitionists (or constructivists, I shall use these terms synonymously) that the basic mathematical notions, above all the notion of function, ought to be interpreted in such a way that the cleavage between mathematics, classical mathematics, that is, and programming that we are witnessing at present disappears. In the case of the mathematical notions of function and set, it is not so much a question of providing them with new meanings as of restoring old ones, whereas the logical notions of proposition, proof, truth etc. are given genuinely new interpretations. It was Brouwer who realized the necessity of so doing: the true source of the uncomputable functions of classical mathematics is not the axiom of choice (which *is* valid intuitionistically) but the law of excluded middle and the law of indirect proof. Had it not been possible to interpret the logical notions in such a way as to validate the axiom of choice, the prospects of constructive mathematics would have been dismal.

The difference, then, between constructive mathematics and programming does not concern the primitive notions of the one or the other, because they are essentially the same, but lies in the programmer's insistence that his programs be written in a formal notation so that they can be read and executed by a machine, whereas, in constructive mathematics as practised by BISHOP (1967), for example, the computational procedures (programs) are normally left implicit in the proofs, so that considerable further work is needed to bring them into a form which makes them fit for mechanical execution.

What I have just said about the close connection between constructive mathematics and programming explains why the intuitionistic type theory (MARTIN-LÖF, 1975), which I began to develop solely with the philosophical motive of clarifying the syntax and semantics of intuitionistic mathematics, may equally well be viewed as a programming language. But for a few concluding remarks, the rest of my talk will be devoted to a fairly complete, albeit condensed, description of this language, emphasizing its character of programming language. As such, it resembles ALGOL 68 and PASCAL in its typing facilities, whereas the way the programs are written and executed makes it more reminiscent of LISP.

The expressions of the theory of types are formed out of variables

x, y, z, \dots

by means of various forms of expression

$$(Fx_1, \dots, x_n)(a_1, \dots, a_m).$$

In an expression of such a form, not all of the variables x_1, \dots, x_n need become bound in all of the parts a_1, \dots, a_m . Thus, for each form of expression, it must be laid down what variables become bound in what parts. For example,

$$\int_a^b f dx$$

is a form of expression $(Ix)(a, b, f)$ with $m = 3$ and $n = 1$ which binds all free occurrences of the single variable x in the third part f . And

$$\frac{df}{dx}(a)$$

is a form of expression $(Dx)(a, f)$ with $m = 2$ and $n = 1$ which binds all free occurrences of the variable x in the second part f .

I shall call an expression, in whatever notation, canonical or normal if it is already fully evaluated, which is the same as to say that it has itself as value. Thus, in decimal arithmetic,

$$0, 1, \dots, 9, 10, 11, \dots$$

are canonical (normal) expressions, whereas

$$2+2, 2\cdot 2, 2^2, 3!, 10^{10^{10}}, \dots$$

are not. An arbitrarily formed expression need not have a value, but, if an expression has a value, then that value is necessarily canonical. This may be expressed by saying that evaluation is idempotent. When you evaluate the value of an expression, you get that value back.

In the theory of types, it depends only on the outermost form of an expression whether it is canonical or not. Thus there are certain forms of expression, which I shall call canonical forms, such that an expression of one of those forms has itself as value, and there are other, noncanonical forms for which it is laid down in some other way how an expression of such a form is evaluated. What I call canonical and noncanonical forms of expression correspond to the constructors and selectors, respectively, of LANDIN (1964). In the context of programming, they might also aptly be called data and program forms, respectively. The table

| Canonical | Noncanonical |
|--------------------------------|--------------------|
| $(\Pi x \in A)B, (\lambda x)b$ | $c(a)$ |
| $(\Sigma x \in A)B, (a, b)$ | $(Ex, y)(c, d)$ |
| $A + B, i(a), j(b)$ | $(Dx, y)(c, d, e)$ |
| $I(A, a, b), r$ | $J(c, d)$ |
| N_0 | $R_0(c)$ |
| $N_1, 0_1$ | $R_1(c, c_0)$ |
| $N_2, 0_2, 1_2$ | $R_2(c, c_0, c_1)$ |
| \dots | \dots |
| $N, 0, a'$ | $(Rx, y)(c, d, e)$ |
| $(Wx \in A)B, \sup(a, b)$ | $(Tx, y, z)(c, d)$ |
| U_0, U_1, \dots | |

displays the primitive forms of expression used in the theory of types, the canonical ones to the left and the noncanonical ones to the right. New primitive forms of expression may of course be added when there is need of them.

The conventions as to what variables become bound in what parts are as follows. Free occurrences of x in B become bound in $(\Pi x \in A)B$, $(\Sigma x \in A)B$ and $(Wx \in A)B$. Free occurrences of x in b become bound in $(\lambda x)b$. Free occurrences of x and y in d become bound in $(Ex, y)(c, d)$. Free occurrences of x in d and y in e become bound in $(Dx, y)(c, d, e)$. Free occurrences of x and y in e become bound in $(Rx, y)(c, d, e)$. And, finally, free occurrences of x, y and z in d become bound in $(Tx, y, z)(c, d)$.

Expressions of the various forms displayed in the table are evaluated according to the following rules. I use

$$b(a_1, \dots, a_n/x_1, \dots, x_n)$$

to denote the result of simultaneously substituting the expressions a_1, \dots, a_n for the variables x_1, \dots, x_n in the expression b . Substitution is the process whereby a program is supplied with its input data, which need not necessarily be in evaluated form.

An expression of canonical form has itself as value. This has already been intimated.

To execute $c(a)$, first execute c . If you get $(\lambda x)b$ as result, then continue

by executing $b(a/x)$. Thus $c(a)$ has value d if c has value $(\lambda x)b$ and $b(a/x)$ has value d .

To execute $(Ex, y)(c, d)$, first execute c . If you get (a, b) as result, then continue by executing $d(a, b/x, y)$. Thus $(Ex, y)(c, d)$ has value e if c has value (a, b) and $d(a, b/x, y)$ has value e .

To execute $(Dx, y)(c, d, e)$, first execute c . If you get $i(a)$ as result, then continue by executing $d(a/x)$. If, on the other hand, you get $j(b)$ as result of executing c , then continue by executing $e(b/y)$ instead. Thus $(Dx, y)(c, d, e)$ has value f if either c has value $i(a)$ and $d(a/x)$ has value f , or c has value $j(b)$ and $e(b/y)$ has value f .

To execute $J(c, d)$, first execute c . If you get r as result, then continue by executing d . Thus $J(c, d)$ has value e if c has value r and d has value e .

To execute $R_n(c, c_0, \dots, c_{n-1})$, first execute c . If you get m_n as result for some $m = 0, \dots, n-1$, then continue by executing c_m . Thus $R_n(c, c_0, \dots, c_{n-1})$ has value d if c has value m_n and c_m has value d for some $m = 0, \dots, n-1$. In particular, $R_0(c)$ has no value. It corresponds to the statement

abort

introduced by DIJKSTRA (1976, p. 26). The pair of forms 0_1 and $R_1(c, c_0)$ together operate in exactly the same way as the pair of forms r and $J(c, d)$. To have them both in the language constitutes a redundancy. $R_2(c, c_0, c_1)$ corresponds to the usual conditional statement

if B then S_1 else S_2

and $R_n(c, c_0, \dots, c_{n-1})$ for arbitrary $n = 0, 1, \dots$ to the statement

with e do $\{c_1 : S_1, \dots, c_n : S_n\}$;

introduced by HOARE (1972, p. 113) and realized by Wirth in PASCAL as the case statement

case e of $c_1 : S_1; \dots; c_n : S_n$ end .

To execute $(Rx, y)(c, d, e)$, first execute c . If you get 0 as result, then continue by executing d . If, on the other hand, you get a' as result, then continue by executing $e(a, (Rx, y)(a, d, e)/x, y)$ instead. Thus $(Rx, y)(c, d, e)$ has value f if either c has value 0 and d has value f , or c has value a' and $e(a, (Rx, y)(a, d, e)/x, y)$ has value f . The closest analogue of the recursion form $(Rx, y)(c, d, e)$ in traditional programming languages is the repetitive statement form

while B do S .

To execute $(Tx, y, z)(c, d)$, first execute c . If you get $\sup(a, b)$ as result, then continue by executing $d(a, b, (\lambda v)(Tx, y, z)(b(v), d)/x, y, z)$. Thus $(Tx, y, z)(c, d)$ has value e if c has value $\sup(a, b)$ and $d(a, b, (\lambda v)(Tx, y, z)(b(v), d)/x, y, z)$ has value e . The transfinite recursion form $(Tx, y, z)(c, d)$ has not yet found any applications in programming. It has, as far as I know, no counterpart in other programming languages.

The traditional way of evaluating an arithmetical expression is to evaluate the parts of the expression before the expression itself is evaluated, as in the computation

$$\begin{array}{r} \overbrace{(3+2)! \cdot 4}^5 \\ \hline \overbrace{120}^{480} \end{array}$$

Thus, traditionally, expressions are evaluated from within, which in programming has come to be known as the applicative order of evaluation. When expressions are evaluated in this way, it is obvious that an expression cannot have a value unless all its parts have values. Moreover, as was explicitly stated as a principle by Frege, the value (Ger. *Bedeutung*) of an expression depends only on the values of its parts. In other words, if a part of an expression is replaced by one which has the same value, the value of the whole expression is left unaffected.

When variable binding forms of expression are introduced, as they are in the theory of types, it is no longer possible, in general, to evaluate the expressions from within. To evaluate $(\lambda x)b$, for example, we would first have to evaluate b . But b cannot be evaluated, in general, until a value has been assigned to the variable x . In the theory of types, this difficulty has been overcome by reversing the order of evaluation: instead of evaluating the expressions from within, they are evaluated from without. This is known as head reduction in combinatory logic and normal order or lazy evaluation in programming. For example, $(\lambda x)b$ is simply assigned itself as value. The term lazy is appropriate since only as few computation steps are performed as are absolutely necessary to bring an expression into canonical form. However, what turns out to be of no significance, it is no longer the case that an expression cannot have a value unless all its parts have values. For example, a' has itself as value even if a has no value. What is significant, though, is that the principle of Frege's referred to above, namely that the value of an expression depends only on the values of its

parts, is irretrievably lost. To make the language work in spite of this loss has been one of the most serious difficulties in the design of the theory of types.

So far, I have merely displayed the various forms of expression used in the theory of types and explained how expressions composed out of those forms are evaluated. The inferential or, as one says in combinatory logic, illative part of the language consists of rules for making judgments of the four forms

A is a type,

A and B are equal types,

a is an object of type A ,

a and b are equal objects of type A ,

abbreviated

A type

$A = B$,

$a \in A$,

$a = b \in A$,

respectively. A judgment of any one of these forms is in general hypothetical, that is, made under assumptions or, to use the terminology of AUTOMATH (DE BRUIJN, 1970), in a context

$$x_1 \in A_1, \dots, x_n \in A_n.$$

In such a context, it is always the case that A_1 is a type, ..., A_n is a type under the preceding assumptions $x_1 \in A_1, \dots, x_{n-1} \in A_{n-1}$. When there is need to indicate explicitly the assumptions of a hypothetical judgment, it will be written

$$A \text{ type } (x_1 \in A_1, \dots, x_n \in A_n),$$

$$A = B(x_1 \in A_1, \dots, x_n \in A_n),$$

$$a \in A(x_1 \in A_1, \dots, x_n \in A_n),$$

$$a = b \in A(x_1 \in A_1, \dots, x_n \in A_n).$$

These, then, are the full forms of judgment of the theory of types.

The first form of judgment admits not only the readings

$$A \text{ is a type (set)},$$

$$A \text{ is a proposition},$$

but also, and this is the reading which is most natural when the language is thought of as a programming language,

A is a problem (task).

Correlatively, the third form of judgment may be read not only

a is an object of type (element of the set) A ,

a is a proof of the proposition A ,

but also

a is a program for the problem (task) A .

The equivalence of the first two readings is the by now well-known correspondence between propositions and types discovered by CURRY (1958, pp. 312–315) and HOWARD (1969), whereas the transition from the second to the third is the KOLMOGOROV (1932) interpretation of propositions as problems or tasks (Ger. *Aufgabe*).

The four forms of judgment used in the theory of types should be compared with the three forms of judgment used (although usually not so called) in standard presentations of first order predicate calculus, whether classical or intuitionistic, namely

A is a formula,

A is true,

a is an individual term.

The first of these corresponds to the form A is a type (proposition), the second is obtained from the form a is an object of type (a proof of the proposition) A by suppressing a , and the third is again obtained from the form a is an object of type A , this time by choosing for A the type of individuals.

In explaining what a judgment of one of the above four forms means, I shall first limit myself to assumption free judgments. Once it has been explained what meanings they carry, the explanations can readily be extended so as to cover hypothetical judgments as well.

A canonical type A is defined by prescribing how a canonical object of type A is formed as well as how two equal canonical objects of type A are formed. There is no limitation on this prescription except that the relation of equality which it defines between canonical objects of type A must be reflexive, symmetric and transitive. If the rules for forming canonical objects

as well as equal canonical objects of a certain type are called the introduction rules for that type, we may thus say with GENTZEN (1934) that a canonical type (proposition) is defined by its introduction rules. For noncanonical A , a judgment of the form

$$A \text{ is a type}$$

means that A has a canonical type as value.

Two canonical types A and B are equal if a canonical object of type A is also a canonical object of type B and, moreover, equal canonical objects of type A are also equal canonical objects of type B , and vice versa. For arbitrary (not necessarily canonical) types A and B , a judgment of the form

$$A = B$$

means that A and B have equal canonical types as values. This finishes the explanations of what a type is and what it means for two types to be equal.

Let A be a type. Remember that this means that A denotes a canonical type, that is, has a canonical type as value. Then a judgment of the form

$$a \in A$$

means that a has a canonical object of the canonical type denoted by A as value. Of course, this explanation is not comprehensible unless we know that A has a canonical type as value as well as what a canonical object of that type is. But we do know this because of the presupposition that A is a type: it is part of the definition of a canonical type how a canonical object of that type is formed, and hence we cannot know a canonical type without knowing what a canonical object of that type is.

Let A be a type and a and b objects of type A . Then a judgment of the form

$$a = b \in A$$

means that a and b have equal canonical objects of the canonical type denoted by A as values. This explanation makes sense since A was presupposed to be a type, that is, to have a canonical type as value, and it is a part of the definition of a canonical type how equal canonical objects of that type are formed.

These meaning explanations are extended to hypothetical judgments by an induction on the number of assumptions. Let it be given as premises for all of the following four explanations that $x_1 \in A_1, \dots, x_n \in A_n$ is a con-

text, that is, that A_1 is a type, ..., A_n is a type under the assumptions $x_1 \in A_1, \dots, x_{n-1} \in A_{n-1}$. By induction hypothesis, we know what this means.

A judgment of the form

$$A \text{ type } (x_1 \in A_1, \dots, x_n \in A_n)$$

means that

$$A(a_1, \dots, a_n/x_1, \dots, x_n) \text{ type}$$

provided

$$\begin{aligned} a_1 &\in A_1, \\ &\vdots \\ a_n &\in A_n(a_1, \dots, a_{n-1}/x_1, \dots, x_{n-1}), \end{aligned}$$

and, moreover,

$$A(a_1, \dots, a_n/x_1, \dots, x_n) = A(b_1, \dots, b_n/x_1, \dots, x_n)$$

provided

$$\begin{aligned} a_1 &= b_1 \in A_1, \\ &\vdots \\ a_n &= b_n \in A_n(a_1, \dots, a_{n-1}/x_1, \dots, x_{n-1}). \end{aligned}$$

Thus it is in the nature of a family of types (propositional function) to be extensional in the sense just described.

Suppose that A and B are types under the assumptions $x_1 \in A_1, \dots, x_n \in A_n$. Then

$$A = B(x_1 \in A_1, \dots, x_n \in A_n)$$

means that

$$A(a_1, \dots, a_n/x_1, \dots, x_n) = B(a_1, \dots, a_n/x_1, \dots, x_n)$$

provided

$$\begin{aligned} a_1 &\in A_1, \\ &\vdots \\ a_n &\in A_n(a_1, \dots, a_{n-1}/x_1, \dots, x_{n-1}). \end{aligned}$$

From this definition, the extensionality of a family of types and the evident transitivity of equality between types, it follows as well that

$$A(a_1, \dots, a_n/x_1, \dots, x_n) = B(b_1, \dots, b_n/x_1, \dots, x_n)$$

provided

$$\begin{aligned} a_1 &= b_1 \in A_1, \\ &\vdots \\ a_n &= b_n \in A_n(a_1, \dots, a_{n-1}/x_1, \dots, x_{n-1}). \end{aligned}$$

Let A be a type under the assumptions $x_1 \in A_1, \dots, x_n \in A_n$. Then

$$a \in A(x_1 \in A_1, \dots, x_n \in A_n)$$

means that

$$a(a_1, \dots, a_n/x_1, \dots, x_n) \in A(a_1, \dots, a_n/x_1, \dots, x_n)$$

provided

$$\begin{aligned} a_1 &\in A_1, \\ &\vdots \\ a_n &\in A_n(a_1, \dots, a_{n-1}/x_1, \dots, x_{n-1}), \end{aligned}$$

and, moreover,

$$a(a_1, \dots, a_n/x_1, \dots, x_n) = b(b_1, \dots, b_n/x_1, \dots, x_n) \in A(a_1, \dots, a_n/x_1, \dots, x_n)$$

provided

$$\begin{aligned} a_1 &= b_1 \in A_1, \\ &\vdots \\ a_n &= b_n \in A_n(a_1, \dots, a_{n-1}/x_1, \dots, x_{n-1}). \end{aligned}$$

Thus, just as in the case of a family of types, it is in the nature of a function to be extensional in the sense of yielding equal objects of the range type when equal objects of the domain types are substituted for the variables of which it is a function.

Let A be a type and a and b objects of type A under the assumptions $x_1 \in A_1, \dots, x_n \in A_n$. Then

$$a = b \in A(x_1 \in A_1, \dots, x_n \in A_n)$$

means that

$$a(a_1, \dots, a_n/x_1, \dots, x_n) = b(a_1, \dots, a_n/x_1, \dots, x_n) \in A(a_1, \dots, a_n/x_1, \dots, x_n)$$

provided

$$\begin{aligned} a_1 &\in A_1, \\ &\vdots \\ a_n &\in A_n(a_1, \dots, a_{n-1}/x_1, \dots, x_{n-1}). \end{aligned}$$

Again, from this definition, the extensionality of a function and the transitivity of equality between objects of whatever type, there follows the stronger property that

$$a(a_1, \dots, a_n/x_1, \dots, x_n) = b(b_1, \dots, b_n/x_1, \dots, x_n) \in A(a_1, \dots, a_n/x_1, \dots, x_n)$$

provided

$$\begin{aligned} a_1 &= b_1 \in A_1, \\ &\vdots \\ a_n &= b_n \in A_n(a_1, \dots, a_{n-1}/x_1, \dots, x_{n-1}). \end{aligned}$$

This finishes my explanations of what judgments of the four forms used in the theory of types mean in the presence of assumptions.

Now to the rules of inference or proof rules, as they are called in programming. They will be presented in natural deduction style, suppressing as usual all assumptions other than those that are discharged by an inference of the particular form under consideration. Moreover, in those rules whose conclusion has one of the forms $a \in A$ and $a = b \in A$, only those premises will be explicitly shown which have these very same forms. This is in agreement with the practice of writing, say, the rules of disjunction introduction in predicate calculus simply

$$\frac{\begin{array}{c} A \text{ true} \\[1ex] B \text{ true} \end{array}}{A \vee B \text{ true}} \quad \frac{}{A \vee B \text{ true}}$$

without showing explicitly the premises that A and B are formulas. For each of the rules of inference, the reader is asked to try to make the conclusion evident to himself on the presupposition that he knows the premises. This does not mean that further verbal explanations are of no help in bringing about an understanding of the rules, only that this is not the place for such detailed explanations. But there are also certain limits to what verbal explanations can do when it comes to justifying axioms and rules of inference. In the end, everybody must understand for himself.

GENERAL RULES

Reflexivity

$$\frac{a \in A}{a = a \in A} \qquad \frac{A \text{ type}}{A = A}$$

Symmetry

$$\frac{a = b \in A}{b = a \in A} \qquad \frac{A = B}{B = A}$$

Pransitivity

$$\frac{a = b \in A \quad b = c \in A}{a = c \in A}$$

$$\frac{A = B \quad B = C}{A = C}$$

Equality of types

$$\frac{a \in A \quad A = B}{a \in B}$$

$$\frac{a = b \in A \quad A = B}{a = b \in B}$$

Substitution

$$\frac{a \in A \quad B \text{ type} \quad (x \in A)}{B(a/x) \text{ type}}$$

$$\frac{a = c \in A \quad B = D \quad (x \in A)}{B(a/x) = D(c/x)}$$

$$\frac{a \in A \quad b \in B \quad (x \in A)}{b(a/x) \in B(a/x)}$$

$$\frac{a = c \in A \quad b = d \in B \quad (x \in A)}{b(a/x) = d(c/x) \in B(a/x)}$$

Assumption

$$x \in A$$

CARTESIAN PRODUCT OF A FAMILY OF TYPES

Π -formation

$$\frac{A \text{ type} \quad B \text{ type} \quad (x \in A)}{(\Pi x \in A)B \text{ type}}$$

$$\frac{A = C \quad B = D \quad (x \in A)}{(\Pi x \in A)B = (\Pi x \in C)D}$$

Π -introduction

$$\frac{(x \in A) \quad b \in B}{(\lambda x)b \in (\Pi x \in A)B}$$

$$\frac{(x \in A) \quad b = d \in B}{(\lambda x)b = (\lambda x)d \in (\Pi x \in A)B}$$

Π -elimination

$$\frac{c \in (\Pi x \in A)B \quad a \in A}{c(a) \in B(a/x)}$$

$$\frac{c = f \in (\Pi x \in A)B \quad a = d \in A}{c(a) = f(d) \in B(a/x)}$$

Π -equality

$$\frac{\begin{array}{c} a \in A \\[1ex] ((\lambda x)b)(a) = b(a/x) \in B(a/x) \end{array}}{(x \in A)} \quad \frac{b \in B}{\lambda x(c(x)) = c \in (\Pi x \in A)B} \quad \frac{c \in (\Pi x \in A)B}{(\lambda x)(c(x)) = c \in (\Pi x \in A)B}$$

DISJOINT UNION OF A FAMILY OF TYPES

 Σ -formation

$$\frac{\begin{array}{c} (x \in A) \\[1ex] A \text{ type} \quad B \text{ type} \end{array}}{(\Sigma x \in A)B \text{ type}} \quad \frac{\begin{array}{c} (x \in A) \\[1ex] A = C \quad B = D \end{array}}{(\Sigma x \in A)B = (\Sigma x \in C)D}$$

 Σ -introduction

$$\frac{\begin{array}{c} a \in A \quad b \in B(a/x) \\[1ex] (a, b) \in (\Sigma x \in A)B \end{array}}{(a, b) \in (\Sigma x \in A)B} \quad \frac{\begin{array}{c} a = c \in A \quad b = d \in B(a/x) \\[1ex] (a, b) = (c, d) \in (\Sigma x \in A)B \end{array}}{(a, b) = (c, d) \in (\Sigma x \in A)B}$$

 Σ -elimination

$$\frac{\begin{array}{c} (x \in A, y \in B) \\[1ex] c \in (\Sigma x \in A)B \quad d \in C((x, y)/z) \end{array}}{(Ex, y)(c, d) \in C(c/z)} \quad \frac{\begin{array}{c} (x \in A, y \in B) \\[1ex] c = e \in (\Sigma x \in A)B \quad d = f \in C((x, y)/z) \end{array}}{(Ex, y)(c, d) = (Ex, y)(e, f) \in C(c/z)}$$

 Σ -equality

$$\frac{\begin{array}{c} (x \in A, y \in B) \\[1ex] a \in A \quad b \in B(a/x) \quad d \in C((x, y)/z) \end{array}}{(Ex, y)((a, b), d) = d(a, b/x, y) \in C((a, b)/z)}$$

DISJOINT UNION OF TWO TYPES

 $+$ -formation

$$\frac{\begin{array}{c} A \text{ type} \quad B \text{ type} \\[1ex] A + B \text{ type} \end{array}}{A + B = C + D} \quad \frac{A = C \quad B = D}{A + B = C + D}$$

+ -introduction

$$\frac{\begin{array}{c} a \in A \\ i(a) \in A+B \end{array}}{j(b) \in A+B} \quad \frac{\begin{array}{c} a = c \in A \\ i(a) = i(c) \in A+B \end{array}}{b = d \in B} \quad \frac{\begin{array}{c} b = d \in B \\ j(b) = j(d) \in A+B \end{array}}$$

+ -elimination

$$\frac{\begin{array}{c} (x \in A) \\ c \in A+B \quad d \in C(i(x)/z) \quad e \in C(j(y)/z) \end{array}}{(Dx, y)(c, d, e) \in C(c/z)} \quad \frac{\begin{array}{c} (x \in A) \\ c = f \in A+B \quad d = g \in C(i(x)/z) \quad e = h \in C(j(y)/z) \end{array}}{(Dx, y)(c, d, e) = (Dx, y)(f, g, h) \in C(c/z)}$$

+ -equality

$$\frac{\begin{array}{c} (x \in A) \\ a \in A \quad d \in C(i(x)/z) \quad e \in C(j(y)/z) \end{array}}{(Dx, y)(i(a), d, e) = d(a/x) \in C(i(a)/z)} \quad \frac{\begin{array}{c} (x \in A) \\ b \in B \quad d \in C(i(x)/z) \quad e \in C(j(y)/z) \end{array}}{(Dx, y)(j(b), d, e) = e(b/y) \in C(j(b)/z)}$$

IDENTITY RELATION

I-formation

$$\frac{\begin{array}{c} A \text{ type} \quad a \in A \quad b \in A \\ I(A, a, b) \text{ type} \end{array}}{I(A, a, b) = I(C, c, d)}$$

I-introduction

$$\frac{a = b \in A}{r \in I(A, a, b)} \quad \frac{a = b \in A}{r = r \in I(A, a, b)}$$

I-elimination

$$\frac{c \in I(A, a, b)}{a = b \in A} \quad \frac{\begin{array}{c} c \in I(A, a, b) \\ d \in C(r/z) \end{array}}{J(c, d) \in C(c/z)} \quad \frac{\begin{array}{c} c = e \in I(A, a, b) \\ d = f \in C(r/z) \end{array}}{J(c, d) = J(e, f) \in C(c/z)}$$

I-equality

$$\frac{a = b \in A \quad d \in C(r/z)}{J(r, d) = d \in C(r/z)}$$

FINITE TYPES

N_n-formation

$$N_n \text{ type} \qquad \qquad N_n = N_n$$

N_n-introduction

$$m_n \in N_n \quad (m = 0, \dots, n-1) \qquad m_n = m_n \in N_n \quad (m = 0, \dots, n-1)$$

N_n-elimination

$$\frac{c \in N_n \quad c_m \in C(m_n/z) \quad (m = 0, \dots, n-1)}{R_n(c, c_0, \dots, c_{n-1}) \in C(c/z)}$$

$$\frac{c = d \in N_n \quad c_m = d_m \in C(m_n/z) \quad (m = 0, \dots, n-1)}{R_n(c, c_0, \dots, c_{n-1}) = R_n(d, d_0, \dots, d_{n-1}) \in C(c/z)}$$

N_n-equality

$$\frac{c_m \in C(m_n/z) \quad (m = 0, \dots, n-1)}{R_n(m_n, c_0, \dots, c_{n-1}) = c_m \in C(m_n/z)} \quad (m = 0, \dots, n-1)$$

NATURAL NUMBERS

N-formation

$$N \text{ type} \qquad \qquad N = N$$

N-introduction

$$0 \in N \qquad \qquad 0 = 0 \in N$$

$$\frac{a \in N}{a' \in N} \qquad \qquad \frac{a = b \in N}{a' = b' \in N}$$

N-elimination

$$\frac{\begin{array}{c} c \in N \quad d \in C(0/z) \quad e \in C(x'/z) \\ \hline (Rx, y)(c, d, e) \in C(c/z) \end{array}}{(x \in N, y \in C(x/z))}$$

$$\frac{\begin{array}{c} c = f \in N \quad d = g \in C(0/z) \quad e = h \in C(x'/z) \\ \hline (Rx, y)(c, d, e) = (Rx, y)(f, g, h) \in C(c/z) \end{array}}{(x \in N, y \in C(x/z))}$$

N-equality

$$\frac{\begin{array}{c} d \in C(0/z) \quad e \in C(x'/z) \\ \hline (Rx, y)(0, d, e) = d \in C(0/z) \end{array}}{(x \in N, y \in C(x/z))}$$

$$\frac{\begin{array}{ccc} a \in N & d \in C(0/z) & e \in C(x'/z) \\ \hline (Rx, y)(a', d, e) = e(a, (Rx, y)(a, d, e)/x, y) \in C(a'/z) \end{array}}{(x \in N, y \in C(x/z))}$$

WELLORDERINGS

W-formation

$$\frac{\begin{array}{c} (x \in A) \\ A \text{ type} \quad B \text{ type} \\ \hline (Wx \in A)B \text{ type} \end{array}}{(x \in A)}$$

$$\frac{\begin{array}{c} (x \in A) \\ A = C \quad B = D \\ \hline (Wx \in A)B = (Wx \in C)D \end{array}}{(x \in A)}$$

W-introduction

$$\frac{a \in A \quad b \in B(a/x) \rightarrow (Wx \in A)B}{\sup(a, b) \in (Wx \in A)B}$$

$$\frac{a = c \in A \quad b = d \in B(a/x) \rightarrow (Wx \in A)B}{\sup(a, b) = \sup(c, d) \in (Wx \in A)B}$$

W-elimination

$$\frac{\begin{array}{c} (x \in A, y \in B \rightarrow (Wx \in A)B, z \in (\Pi v \in B)C(y(v)/w)) \\ c \in (Wx \in A)B \qquad \qquad \qquad d \in C(\sup(x, y)/w) \end{array}}{(Tx, y, z)(c, d) \in C(c/w)}$$

$$\frac{\begin{array}{c} (x \in A, y \in B \rightarrow (Wx \in A)B, z \in (\Pi v \in B)C(y(v)/w)) \\ c = e \in (Wx \in A)B \qquad \qquad \qquad d = f \in C(\sup(x, y)/w) \end{array}}{(Tx, y, z)(c, d) = (Tx, y, z)(e, f) \in C(c/w)}$$

W-equality

$$\frac{\begin{array}{c} (x \in A, y \in B \rightarrow (Wx \in A)B, z \in (\Pi v \in B)C(y(v)/w)) \\ a \in A \qquad b \in B(a/x) \rightarrow (Wx \in A)B \qquad d \in C(\sup(x, y)/w) \\ (Tx, y, z)(\sup(a, b), d) = d(a, b, (\lambda v)(Tx, y, z)(b(v), d)/x, y, z) \end{array}}{\in C(\sup(a, b)/w)}$$

UNIVERSES

U_n-formation

$$U_n \text{ type} \qquad \qquad U_n = U_n$$

U_n-introduction

$$\frac{\begin{array}{c} (x \in A) \\ A \in U_n \quad B \in U_n \\ (\Pi x \in A)B \in U_n \end{array}}{(\Sigma x \in A)B \in U_n} \qquad \qquad \frac{\begin{array}{c} (x \in A) \\ A = C \in U_n \quad B = D \in U_n \\ (\Pi x \in A)B = (\Pi x \in C)D \in U_n \end{array}}{D \in U_n}$$

$$\frac{\begin{array}{c} (x \in A) \\ A \in U_n \quad B \in U_n \\ (\Sigma x \in A)B \in U_n \end{array}}{A + B \in U_n} \qquad \qquad \frac{\begin{array}{c} (x \in A) \\ A = C \in U_n \quad B = D \in U_n \\ (\Sigma x \in A)B = (\Sigma x \in C)D \in U_n \end{array}}{C + D \in U_n}$$

$$\frac{\begin{array}{c} A \in U_n \quad a \in A \quad b \in A \\ I(A, a, b) \in U_n \end{array}}{I(A, a, b) \in U_n} \qquad \qquad \frac{\begin{array}{c} A = C \in U_n \quad a = c \in A \quad b = d \in A \\ I(A, a, b) = I(C, c, d) \in U_n \end{array}}{I(C, c, d) \in U_n}$$

$$\begin{array}{ll}
 N_0 \in U_n & N_0 = N_0 \in U_n \\
 N_1 \in U_n & N_1 = N_1 \in U_n \\
 \vdots & \vdots \\
 N \in U_n & N = N \in U_n \\
 \\
 \frac{(x \in A)}{A \in U_n \quad B \in U_n} & \frac{(x \in A)}{A = C \in U_n \quad B = D \in U_n} \\
 \frac{(Wx \in A)B \in U_n}{U_0 \in U_n} & \frac{(Wx \in A)B = (Wx \in C)D \in U_n}{U_0 = U_0 \in U_n} \\
 \\
 \frac{\vdots}{U_{n-1} \in U_n} & \frac{\vdots}{U_{n-1} = U_{n-1} \in U_n}
 \end{array}$$

U_n-elimination

$$\begin{array}{ll}
 \frac{A \in U_n}{A \text{ type}} & \frac{A = B \in U_n}{A = B} \\
 \\
 \frac{A \in U_n}{A \in U_{n+1}} & \frac{A = B \in U_n}{A = B \in U_{n+1}}
 \end{array}$$

An example will demonstrate how the language works. Let the premises

$$\begin{aligned}
 & A \text{ type,} \\
 & B \text{ type } (x \in A), \\
 & C \text{ type } (x \in A, y \in B)
 \end{aligned}$$

be given. Make the abbreviation

$$\frac{(\Pi x \in A)B}{A \rightarrow B}$$

provided the variable x does not occur free in B . Then

$$(\Pi x \in A)(\Sigma y \in B)C \rightarrow (\Sigma f \in (\Pi x \in A)B)(\Pi x \in A)C(f(x)/y)$$

is a type which, when read as a proposition, expresses the axiom of choice. I shall construct an object of this type, an object which may at the same time be interpreted as a proof of the axiom of choice. Assume

$$x \in A, \quad z \in (\Pi x \in A)(\Sigma y \in B)C.$$

By Π -elimination,

$$z(x) \in (\Sigma y \in B)C.$$

Make the abbreviations

$$\frac{(Ex, y)(c, x),}{p(c)} \quad \frac{(Ex, y)(c, y).}{q(c)}$$

By Σ -elimination,

$$\begin{aligned} p(z(x)) &\in B, \\ q(z(x)) &\in C(p(z(x))/y). \end{aligned}$$

By Π -introduction,

$$(\lambda x)p(z(x)) \in (\Pi x \in A)B,$$

and, by Π -equality,

$$((\lambda x)p(z(x)))(x) = p(z(x)) \in B.$$

By symmetry,

$$p(z(x)) = ((\lambda x)p(z(x)))(x) \in B,$$

and, by substitution,

$$C(p(z(x))/y) = C(((\lambda x)p(z(x)))(x)/y).$$

By equality of types,

$$q(z(x)) \in C(((\lambda x)p(z(x)))(x)/y),$$

and, by Π -introduction,

$$(\lambda x)q(z(x)) \in (\Pi x \in A)C(((\lambda x)p(z(x)))(x)/y).$$

By Σ -introduction,

$$((\lambda x)p(z(x)), (\lambda x)q(z(x))) \in (\Sigma f \in (\Pi x \in A)B)(\Pi x \in A)C(f(x)/y).$$

Finally, by Π -introduction,

$$\begin{aligned} &(\lambda z)((\lambda x)p(z(x)), (\lambda x)q(z(x))) \\ &\in (\Pi x \in A)(\Sigma y \in B)C \rightarrow (\Sigma f \in (\Pi x \in A)B)(\Pi x \in A)C(f(x)/y). \end{aligned}$$

Thus

$$(\lambda z)((\lambda x)p(z(x)), (\lambda x)q(z(x)))$$

is the sought for proof of the axiom of choice.

To conclude, relating constructive mathematics to computer programming seems to me to have a beneficial influence on both parties. Among the benefits to be derived by constructive mathematics from its association with computer programming, one is that you see immediately why you cannot rely upon the law of excluded middle: its uninhibited use would

lead to programs which you did not know how to execute. Another is that you see the point of introducing a formal notation not only for propositions, as in propositional and predicate logic, but also for their proofs: this is necessary in order to make the methods of computation implicit in intuitionistic (constructive) proofs fit for automatic execution. And a third is that you see the point of formalizing the process of reasoning: this is necessary in order to have the possibility of automatically verifying the programs' correctness. In fact, if the AUTOMATH proof checker had been written for the theory of types instead of the language AUTOMATH, we would already have a language with the facility of automatic checking of the correctness of the programs formed according to its rules.

In the other direction, by choosing to program in a formal language for constructive mathematics, like the theory of types, one gets access to the whole conceptual apparatus of pure mathematics, neglecting those parts that depend critically on the law of excluded middle, whereas even the best high level programming languages so far designed are wholly inadequate as mathematical languages (and, of course, nobody has claimed them to be so). In fact, I do not think that the search for logically ever more satisfactory high level programming languages can stop short of anything but a language in which (constructive) mathematics can be adequately expressed.

References

- BISHOP, 1967, *Foundations of constructive analysis* (McGraw-Hill, New York)
- DE BRUIN, N. G., 1970, *The mathematical language AUTOMATH, its usage, and some of its extensions*, in: Symposium on Automatic Demonstration, Lecture Notes in Mathematics, vol. 125 (Springer-Verlag, Berlin), pp. 29–61
- CURRY, H. B., 1958, *Combinatory logic*, vol. I (North-Holland, Amsterdam)
- DAHL, O.-J., E. W. DIJKSTRA, and C. A. R. HOARE, 1972, *Structured programming* (Academic Press, London)
- DIJKSTRA, E. W., 1976, *A discipline of programming* (Prentice-Hall, Englewood Cliffs, N. J.)
- GENTZEN, G., 1934, *Untersuchungen über das logische Schliessen*, Mathematische Zeitschrift, vol. 39, pp. 176–210, 405–431
- HOARE, C. A. R., 1972, *Notes on data structuring*, in: DAHL, DIJKSTRA and HOARE (1972), pp. 83–174
- HOWARD, W. A., 1969, *The formulae-as-types notion of construction*
- KOLMOGOROV, A. N., 1932, *Zur Deutung der intuitionistischen Logik*, Mathematische Zeitschrift, vol. 35, pp. 58–65
- LANDIN, 1964, *The mechanical evaluation of expressions*, Computer Journal, vol. 6, pp. 308–320
- MARTIN-LÖF, P., 1975, *An intuitionistic theory of types: predicative part*, in: Logic Colloquium '73, eds. H. E. Rose and J. C. Shepherdson (North-Holland, Amsterdam), pp. 73–118
- WIRTH, N., 1971, *The programming language Pascal*, Acta Informatica, vol. 1, pp. 35–63

CRUMBLY SPACES

YURI GUREVICH

Ben-Gurion University, Beer-Sheva, Israel, and Simon Fraser University, Canada

1. Introduction

HENSON, JOCKUSCH, JR., RUBEL, TAKEUTI (1977) define the first-order theory of a top. space U as the first-order theory of the lattice of closed subsets of U . We define and study crumbly spaces in order to answer the following question of HENSON *et al.* (1977).

Q3. Are any two 0-dimensional separable metric spaces without isolated points elementary equivalent? In particular, are the rationals, the irrationals and the Cantor set elementary equivalent as top. spaces?

The answer is yes. All crumbly spaces without isolated points have the same first-order theory, and any 0-dimensional separable metric space is crumbly. The first-order theory of crumbly spaces is decidable. The method used is that of GUREVICH (1979), essentially Shelah's.

Speaking about a space or a top. space, we always mean a non-empty T_1 space. A space and its universe are denoted in the same way. The derivative (the set of limit points) of a point set X is denoted by X' . As usual, X is a deleted neighbourhood of a point y if $y \notin X$ and $X \cup \{y\}$ is a nbd (neighbourhood) of y .

Considering a chain (linearly ordered set) as a top. (= topological) space, we always have in mind the interval topology.

“W.l.o.g.” and “nwd” abbreviate “without loss of generality” and “nowhere dense”, respectively.

2. Congruences

Let U be a top. space and let E be an equivalence relation on U with closed fibers (= equivalence classes). The fibers of E form the quotient space U/E . By definition a set A of fibers is open in U/E if $\bigcup A$ is open in U .

E will be called a *congruence* if

- (Co1) All non-singleton fibers of E are open, and
- (Co2) The quotient mapping $x \rightarrow x/E$ is closed.

The quotient mapping is open if (Co1) holds. However, (Co1) does not imply (Co2). Consider a chain

$$\dots -2, -1, 0, 1, 2, \dots, \omega$$

and the equivalence relation with fibers $\{0\}, \{-1, +1\}, \{-2, +2\}, \dots, \{\omega\}$.

For $X \subseteq U$ let $X_E = \bigcup \{x/E : x \in X\}$.

CLAIM 1. *In presence of (Co1) condition (Co2) is equivalent to*

(Co3) *If $X \subseteq U$ and $y \in (X_E)' - X_E$, then each nbd of y includes some x/E with $x \in X$.*

PROOF: (Co2) means that X_E is closed for any closed X . For a closed X (Co3) implies $(X_E)' \subseteq X_E \cup X' = X_E$. If X, y and a nbd H of y give a counterexample for (Co3) select a point fx from each $(x/E) - H$ with $x \in X$. Let F be the closure of $\{fx : x \in X\}$. Then $X_E \subseteq F_E$ but $y \notin F_E$; hence F_E is not closed and (Co2) fails.

CLAIM 2. *Suppose that E is a congruence on U . The reduction of E on an arbitrary subspace V of U is a congruence on V .*

PROOF: Check (Co1) and (Co3).

CLAIM 3. *Let U be a chain.*

(i) *If all non-singleton fibers of E are clopen and convex, then E is a congruence.*

(ii) *Suppose that E is a congruence on U . Define $x \sim y$ if $x E y$ and $x E z$ for every z between x and y . Then \sim is a congruence on U .*

PROOF: (i) Check condition (Co3). (ii) Use statement (i).

Given a top. space I and a disjoint family $\{V_i : i \in I\}$ of top. spaces such that V_i is singleton for $i \in I'$, we define $V = \sum \{V_i : i \in I\}$ as follows. $x \in V$ if x belongs to some V_i and an arbitrary $X \subseteq V$ is closed in V if $V_i \cap X$ is closed in V_i , for each i and $\{i : V_i \text{ meets } X\}$ is closed in I (so that X is open in V iff $V_i \cap X$ is open in V_i for each i and $\{i : V_i \subseteq X\}$ is open in I). Check that V is really a top. space, every V_i is a closed subspace of V and the relation “ x, y are in the same V_i ” is a congruence on V . We say that V is the *discrete sum* of spaces V_i if I is discrete.

CLAIM 4. Suppose that E is a congruence on U . Form $I \subseteq U$ by selecting a point from each fiber of E . Then I is closed, $U = \sum \{i/E: i \in I\}$ and the mapping $i/E \rightarrow i$ is a homeomorphism of U/E onto I .

PROOF: If $x \in I'$, then x/E cannot be open; hence x/E is singleton and $x \in I$.

Given $X \subseteq U$ with $(i/E) \cap X$ closed for each i , we check that X is closed iff $J = \{i: i/E \text{ meets } X\}$ is closed. If X is closed and $y \in J'$, then $y \in (X_E)' \subseteq X_E$ and y/E is singleton; hence $y \in X$. If J is closed and $y \in J'$, then $y \in J'_E \subseteq J_E$ and either $y \in ((y/E) \cap X)' \subseteq X$ or $y/E = \{y\} \subseteq X$.

If A is a closed set in U/E , then $\bigcup A$ is closed in U and $(\bigcup A) \cap I$ is closed in I . If J is closed in I , it is closed in U ; hence $\{i/E: i \in J\}$ is closed in U/E .

3. Crumbly spaces

Given a family A of open sets in a top. space U and a congruence E on U , we say that E crumbles A if E is the identity on $U - \bigcup A$ and for each $x \in \bigcup A$ the fiber x/E is an open subset of some $a \in A$. If E crumbles $A = \{G\}$, we say that E crumbles G .

DEFINITION. A space U is *crumbly* if

(Cr1) For each family A of open sets in U there is a congruence on U crumbling A , and

(Cr2) Each discrete $X \subseteq U$ can be split into Y and $X - Y$ such that $Y' = (X - Y)'$.

(Cr1) does not imply (Cr2): consider $U = \omega \cup \{F\}$ where F is a non-principal ultrafilter on ω and $X \subseteq U$ is open iff $X \subseteq \omega$ or $X - \{F\} \in F$.

CLAIM 1. Let U be a crumbly space.

(i) Any subspace of U is crumbly.

(ii) For any congruence E the quotient space U/E is crumbly.

PROOF: (i) is straightforward. (ii) follows from (i) and Claim 4 in § 2.

LEMMA 2. Let f be a continuous mapping from a space U onto a space I and $U_i = f^{-1}(i)$ for $i \in I$. Suppose that if $y \in U_i \cap (U - U_i)'$ and H is a nbd of y , then there is a deleted nbd J of i in I such that $\bigcup \{U_j: j \in J\} \subseteq H$. If I and all subspaces U_i satisfy (Cr2), then U satisfies (Cr2).

PROOF: Given a discrete $X \subseteq U$ form $K = \{i: U_i \cap X \text{ is singleton}\}$ and split it into K_1, K_2 such that $K'_1 = K'_2$. Split each $U_i \cap X$ into X_{1i}, X_{2i} such that $X'_{1i} = X'_{2i}$, and both X_{si} are not empty if $|U_i \cap X| \geq 2$, and X_{si} is not empty if $i \in K_s$. Form $X_s = \bigcup \{X_{si}: i \in I\}$.

Given $y \in U_i \cap X'$ and a nbd H of y , we show that H meets X_e . That is clear if $y \in (U_i \cap X)'$. Suppose $y \in (X - U_i)'$. W.l.o.g. $H - U_i = \bigcup \{U_j : j \in J\}$ for some deleted nbd J of i . If $|U_j \cap X| \geq 2$ for some $j \in J$, then H meets X_e . Otherwise J meets K ; hence J meets K_e , hence H meets X_e . \square

THEOREM 3. $U = \sum \{U_i : i \in I\}$ is crumbly if I and all U_i are crumbly spaces.

PROOF: In virtue of Lemma 2 it suffices to check (Cr1). Given a family A of open sets in U , form the family B of open sets $\{i : U_i \subseteq a\}$ in I where $a \in A$ and crumble B by a congruence e on I . If $i \in I - \bigcup B$, crumble $A|U_i$ by a congruence E_i on U_i . For $x \in U_i$ and $y \in U_j$, define xEy if iej or $i = j$ and xE_iy . E is an equivalence relation on U . It crumbles A if it is a congruence. In the remaining part of the proof we show that E satisfies conditions (Co1) and (Co3).

If F is a fiber of E , then either $F = \bigcup \{U_i : i \in J\}$ where J is an open fiber of e or else F is a fiber of some E_i . In the first case F is open in U . In the second case, if F is not singleton, then it is open in U_i and U_i is open in U , hence F is open in U .

Let $X \subseteq U$, $y \in (X_E)' - X_E$, $y \in U_i$ and let H be a nbd of y . W.l.o.g. $i \in I'$ and there is a nbd J of i with $\bigcup \{U_j : j \in J\} \subseteq H$. If $x/E \subseteq U_i$ for some $x \in X$ and $j \in J$, then $x/E \subseteq H$. Otherwise, by (Co3) applied to e , J includes a fiber K of e with $\bigcup \{U_j : j \in K\} = x/E$ for some $x \in X$ which implies $x/E \subseteq H$.

CLAIM 4. Any crumbly space without isolated points has a perfect nwd set.

PROOF: Build a sequence E_0, E_1, \dots of congruences and a sequence X_0, X_1, \dots of point sets in such a way that X_n comprises a point from each non-singleton fiber of E_n , and E_{n+1} refines E_n and crumbles the complement of X_n , and all points are E_0 -equivalent.

For each $n > 0$ choose $Y_n \subset X_n$ such that $Y'_n = (X_n - Y_n)'$. The difference between $\bigcup \{X_n : n \geq 0\}$ and $\bigcup \{(Y_n)_{E_n} : n > 0\}$ is perfect and nwd.

4. Crumbly chains

Let U be a chain.

CLAIM 1. U satisfies (Cr2).

PROOF: Given a discrete $X \subseteq U$ define xey if $x, y \in X$ and there is no points of X' between x, y . Let X_1 be the union of non-singleton fibers of e . Splitting the non-singleton fibers one gets $Y_1 \subseteq X_1$ with $Y'_1 = (X_1 - Y_1)'$. It remains to split $X - X_1$. W.l.o.g. $X_1 = 0$.

Let κ be an infinite cardinal. An interval I of the subchain X will be called κ -good if $\kappa = |I| = |(x, y)|$ for any $x < y$ in I . Any non-empty non-singleton interval of the subchain X has a subinterval κ -good for some κ . If I is κ -good arrange a list $\langle I_\alpha : \alpha < \kappa \rangle$ of all subintervals (x, y) with $x < y$ in I and build disjoint $Y, Z \subseteq I$ meeting each I_α , then $Y' = (I - Y)'$. Now consider the equivalence relation $x E y$ if $x, y \in X$ and either $x = y$ or there is a subset Y of $I = \{z \in X : z \text{ between } x, y\}$ such that $Y' = (I - Y)'$. \square

U is 0-dimensional iff no interval of U with at least two points is connected iff between any two different points of U there is a jump or an empty Dedekind cut. (A jump between $x < y$ means a pair $u < v$ such that $x \leq u < v \leq y$ and the interval (u, v) is empty.) Suppose that U is 0-dimensional.

THEOREM 2. *The following statements are equivalent:*

- (1) *U is crumbly,*
- (2) *Each non-empty open interval of U without a minimum (respectively maximum) point can be partitioned into a chain of open subintervals without a first (respectively last) member.*
- (3) *There are no κ, S and $f: S \rightarrow U$ such that κ is a regular uncountable cardinal, S is a stationary subset of κ , f is a continuous mapping from the subspace S of the chain κ into U preserving or reversing the order.*

PROOF: (1) \rightarrow (2). Given a non-empty open interval I without a maximum point crumble the family of open initial proper subintervals of I and use Claim 3 in § 2.

(2) \rightarrow (3). Let κ, S and $f: S \rightarrow U$ give a counterexample for (3) where f preserves the order. Partition the least initial interval of U including fS into a chain J of open subintervals without a last member. The set $C = \{\alpha : \text{there is an initial proper subinterval } K \text{ of } J \text{ such that for each } \beta \in S, f\beta \in \bigcup K \text{ iff } \beta < \alpha\}$ is closed and unbounded in κ . Consider $\alpha \in C \cap S$ to get a contradiction.

(3) \rightarrow (2). Let I be a non-empty open interval of U without a maximum point and κ be the cofinality of I . Choose an ascending sequence $\langle I_\alpha : \alpha < \kappa \rangle$ of initial subintervals of I covering I and such that no I_α has a supremum in $I - I_\alpha$. If $\kappa = \omega$ partition I in the obvious way. If $\kappa > \omega$ consider the function $f\alpha = \sup \bigcup \{I_\beta : \beta < \alpha\}$. By (3) the domain S of f is not stationary. Use a closed unbounded $C \subseteq \kappa - S$ to partition I .

(2) \rightarrow (1). By Claim 1 it suffices to check (Cr 1). Let A be a family of open sets in U . A clopen interval I will be called a *crumb* if there is $a \in A$ including I . A set $X \subseteq U$ will be called *good* if there is a disjoint family B of crumbs

such that $\bigcup B$ is convex and includes X . It suffices to prove that each maximal interval in $\bigcup A$ is good. W.l.o.g. $\bigcup A = U$.

Define xEy if $\{x, y\}$ is good. If $x < y < z$, xEy and yEz , take disjoint families B , C of crumbs such that $\bigcup B$, $\bigcup C$ are convex and include $\{x, y\}$, $\{y, z\}$, respectively. If $y \in b \in B$ and $y \in c \in C$ then $\{x, z\}$ is covered by $b \cap c$, the part of b below c , the part of c above b , the crumbs of B below b and the crumbs of C above c . Hence E is an equivalence relation.

Any fiber of E is an open interval. It suffices to prove that each fiber of E is good. Now use (2).

COROLLARY 3. *Any separable 0-dimensional metric space is crumbly.*

PROOF: By classical theorems any separable metric space is second-countable, and any 0-dimensional second-countable metric space is embeddable into the Cantor set. The Cantor set is crumbly by Theorem 2. Now use Claim 1 of § 3.

5. The adjusted theory

A point set X is *sandy* if each point of X is isolated in the space.

CLAIM 1. *Let U be a crumbly space with nwd U' . Each closed $X \subseteq U'$ is the derivative of some sandy set.*

PROOF: Crumble $U - X$ by a congruence and pick an isolated point in each open fiber.

CLAIM 2. *Let U be a crumbly space with nwd U' . For each first-order sentence φ in the language $\{\leq\}$ the following statements are equivalent:*

- (1) *The collection of closed subsets of U' ordered by inclusion satisfies φ .*
- (2) *The collection of sandy subsets of U with $X \leq Y$ meaning $X' \subseteq Y'$ satisfies φ .*

PROOF: Identification of indistinguishable members in (2) gives a model isomorphic to the model of (1).

THEOREM 3. *For each crumbly space I there is a crumbly superspace U such that $I = U'$ and I is nwd in U .*

PROOF: The idea is to sew on tails to enough points of I . Choose an everywhere dense $I_0 \subseteq I$ and a disjoint family $\{U_i : i \in I\}$ of crumbly spaces such that $U_i = \{i\}$ if $i \in I - I_0$ and $U'_i = \{i\}$ if $i \in I_0$. Define a space U as

follows: $x \in U$ if x belongs to some U_i , and an arbitrary $X \subseteq U$ is closed in U if $U_i \cap X$ is closed in U_i for each i and X includes the derivative in I of $\{i: U_i \text{ meets } X\}$ (so that X is open in U iff $U_i \cap X$ is open in U_i for each i and $\{i: U_i \subseteq X\}$ includes a deleted nbd in I of any $j \in I \cap X$). Check that U is really a top. space, all spaces U_i are closed subspaces of U , $I = U'$, I is nwd in U , and associating i with each point in U_i gives a continuous mapping from U onto I .

We prove that U is crumbly. In virtue of Lemma 2 in § 3 it suffices to check (Cr1) only. Given a family A of open sets in U crumble $A|I$ by a congruence e on I . If $J \subseteq (\cup A) \cap I$ is a fiber of e , choose $gJ \in A$ with $J \subseteq gJ$ and form $hJ = gJ \cap \cup \{U_i: i \in J\}$. The sets hJ are clopen and disjoint. Consider the equivalence relation E on U whose only non-singleton fibers are the sets hJ . Clearly, E crumbles A if it is a congruence. Clearly, E satisfies (Co1). It remains to check (Co3). Let $X \subseteq U$, $y \in (X_E)' - X_E$ and H be a nbd of y in U . Then $y \in I$ and there is a deleted nbd K of y in I such that $\cup \{U_i: i \in K\} \subseteq H$. If H does not include any singleton x/E with $x \in X$, then, by (Co3) applied to e , K includes some J with $hJ = x/E$ for some $x \in X$ which implies $x/E \subseteq H$.

DEFINITION. The *adjusted first-order theory* of a crumbly space U with a nwd derivative is the first-order theory of the Boolean algebra of sandy subsets of U with an additional relation $X \leq Y$ meaning $X' \subseteq Y'$. T is the adjusted first-order theory of all crumbly spaces with nwd derivatives.

COROLLARY 4. *The first-order theory of crumbly spaces is interpretable in T . All crumbly spaces without isolated points have the same first-order theory if all crumbly spaces U such that U' is nwd and $0 \neq U' = U''$ have the same adjusted first-order theory.*

CLAIM 5. *Every two crumbly spaces U, V with singleton derivative have the same adjusted first-order theory.*

PROOF: We describe a winning strategy for us (the player II) against the devil (the player I) in the Ehrenfeucht game $G_n(U, V)$.

If X_1, \dots, X_m and Y_1, \dots, Y_m were chosen during the first m moves consider all intersections $\pm X_1 \cap \dots \cap \pm X_m$ and $\pm Y_1 \cap \dots \cap \pm Y_m$ where $+Z = Z$ and $-Z$ is the complement of Z in the respective Boolean algebra of sandy sets. It gives A_1, \dots, A_l in U and B_1, \dots, B_l in V , respectively, where $l = 2^m$. Our strategy is to satisfy the following requests: $A'_k = 0$ iff $B'_k = 0$, and either both $|A_k|, |B_k|$ are $\geq 2^{n-k}$ or $|A_k| = |B_k|$.

6. The theory of sum

In order to analize theory T we use the method of GUREVICH (1979).

CLAIM 1. *The universal fragment of T is decidable.*

PROOF: It suffices to prove satisfiability in T of any quantifier-free formula $\varphi \wedge \psi$ where φ states that v_1, \dots, v_l partition 1 and ψ describes a quasi-order on $0, v_1, \dots, v_l$ with $0 \leq v_k$ for each k . Consider the discrete sum $U_1 + \dots + U_l$ where U_k is a copy of ω if $\psi \vdash v_k \leq 0$ and of $\omega + 1$ if $\psi \vdash 0 < v_k$. Choose disjoint infinite sandy U_{k1}, \dots, U_{kl} in each U_k and interprete each v_k as $\bigcup \{U_{jk} : \psi \vdash v_j \leq v_k\}$. \square

A crumbly space U with nwd U' augmented by a sequence $P = \langle P_1, \dots, P_l \rangle$ of sandy subsets of U will be called an *ausp* of weight l . If V is a clopen subspace of U , then $\langle V, P|V \rangle$ will be called a *subausp* of $\langle U, P \rangle$.

The 0 -theory $\text{Th}^0(U, P)$ of an ausp $\langle U, P \rangle$ of weight l is a quantifier-free description of that ausp in variables v_1, \dots, v_l (see a more rigorous definition in GUREVICH, 1979). Let $\xi = \langle \xi_n : n < \omega \rangle$ be a sequence of natural numbers with a tail of zeros. The $n - \xi$ -theory $\text{Th}_\xi^n(U, P)$ of an ausp $\langle U, P \rangle$ is defined by inductions:

$$\begin{aligned}\text{Th}_\xi^0(U, P) &= \text{Th}^0(U, P), \\ \text{Th}_\xi^{n+1}(U, P) &= \{\text{Th}_\xi^n(U, P \hat{Q}) : lh(Q) = \xi_n\}.\end{aligned}$$

The true value of $U \models \varphi(P)$ with elementary φ is computable from an appropriate $\text{Th}_\xi^n(U, P)$ (with n, ξ computable from φ).

Sets $\text{Tr}_\xi^n(l)$ and l -traces are defined by induction. $\text{Tr}^0(l) = \text{Tr}_\xi^0(l) = \{\text{Th}^0(M) : M \text{ is an ausp of weight } l\}$. $s \in \text{Tr}^0(l)$ is an l -trace of $t \in \text{Tr}^0(l+m)$ if there are U, P, Q such that $s = \text{Th}^0(U, P)$ and $t = \text{Th}^0(U, P \hat{Q})$. $\text{Tr}_\xi^{n+1}(l) = \{t \subseteq \text{Tr}_\xi^n(l + \xi_n) : \text{all members of } t \text{ have the same } l\text{-trace}\}$. $s \in \text{Tr}^0(l)$ is an l -trace of $t \in \text{Tr}_\xi^{n+1}(l+m)$ if s is the l -trace of each member of t . Check that $\text{Tr}_\xi^n(l)$ is recursive in n, ξ, l and contains the $n - \xi$ -theory of any ausp of weight l .

Let $\tilde{U} = \langle U, P \rangle$ be an ausp of weight l , let E be a congruence on U , X range over open fibers of E and let \tilde{X} be the subausp $\langle X, P|X \rangle$. We introduce some more notation. For each $t \in \text{Tr}_\xi^n(l)$ the set $\{X : \text{Th}_\xi^n(\tilde{X}) = t\}$ will be denoted by $(\tilde{U}/E)_\xi^n t$. The sequence $\langle (\tilde{U}/E)_\xi^n t : t \in \text{Tr}_\xi^n(l) \rangle$ (we consider $\text{Tr}_\xi^n(l)$ being linearly ordered in a standard way) will be denoted by $(\tilde{U}/E)_\xi^n$. Finally, $[\tilde{U}/E]_\xi^n$ is the ausp $\langle U/E, (\tilde{U}/E)_\xi^n \rangle$. ξ may be omitted everywhere if $n = 0$.

CLAIM 2. $\text{Th}^0(\tilde{U})$ is computable from $\text{Th}^0[\tilde{U}/E]^0$.

PROOF: If τ is a Boolean term in variables v_1, \dots, v_l , then $\tau(P) \cap X = \tau(P|X)$ (check by induction on τ). Given $\text{Th}^0[\tilde{U}/E]^0$, we compute whether $\tau(P) = 0$ and whether $\tau_1(P) \leq \tau_2(P)$.

$\tau(P) = 0$ iff $\tau(P|X) = 0$ for each X iff $(\tilde{U}/E)^0 t \neq 0$ only for t implying $\tau = 0$.

$\tau_1(P) \leq \tau_2(P)$ iff $\tau_1(P|X) \leq \tau_2(P|X)$ for each X and $\{X : \tau_1(P|X) \neq 0\}' \subseteq \{X : \tau_2(P|X) \neq 0\}'$ in U/E iff $(\tilde{U}/E)^0 t \neq 0$ only for t implying $\tau_1 \leq \tau_2$ and

$$\cup \{(\tilde{U}/E)^0 t : t \vdash \tau_1 \neq 0\} \leq \cup \{(\tilde{U}/E)^0 t : t \vdash \tau_2 \neq 0\}.$$

η denotes below a sequence $\langle \eta n : n < \omega \rangle$ of natural numbers with a tail of zeros.

THEOREM 3. There is a recursive function $\eta = T(n, \xi, l)$ such that $\text{Th}_\xi^n(\tilde{U})$ is computable from n, ξ, l and $\text{Th}_\eta^n[\tilde{U}/E]_\xi^n$.

PROOF: See Theorem 2.2 in GUREWICH (1979) and Claim 2.

If V is the discrete sum of spaces V_j , then $\text{Th}_\xi^n(V, Q)$ will be called the *discrete sum* of $\text{Th}_\xi^n(V_j, Q|V_j)$. That defines the commutative semigroup of $n-\xi$ -theories of ausps of weight l . By Theorem 3 the sum $s+t$ of elements of that semigroup is computable from n, ξ, l, s, t .

If I is a crumbly space with nwd I' , $U = \sum \{U_i : i \in I\}$, P is a sequence of l sandy subsets of U and $\text{Th}_\xi^n(U_i, P|U_i) = s$ for each isolated $i \in I$ we write $\text{Th}_\xi^n(U, P) = s \cdot I$. (In particular, $s2 = s+s$.) By Theorem 3, $s \cdot I$ is computable from n, ξ, l, s and the $n-T(n, \xi, l)$ -theory of I .

CLAIM 4. Let S be the semigroup of $n-\xi$ -theories of ausps of weight l .

- (i) There is an integer m such that $tm = t\omega = t \cdot I$ for every $t \in S$ and every infinite discrete space I .
- (ii) The semigroup generated by a non-empty subset S_0 of S is the closure of S_0 under discrete sums.
- (iii) $t(\omega+1) = t \cdot I$ for any crumbly space I with singleton derivative.

PROOF: (i) is clear. (ii) follows from (i). To prove (iii) use Claim 5 in § 5.

7. Uniformity

Two ausps are 0-equivalent ($n-\xi$ -equivalent) if they have the same 0-theory ($n-\xi$ -theory). An ausp $\tilde{U} = \langle U, P \rangle$ is 0-uniform if $0 \neq U' = U''$, and either P is the empty sequence or P partitions 1, and $U' = P'_i$ for each non-empty member P_i of P . Check that all non-sandy subausps of a 0-uni-

form ausp are 0-equivalent. \tilde{U} is $n-\xi$ -uniform if all non-sandy subausps of \tilde{U} are $n-\xi$ -equivalent. Below $\tilde{U} = \langle U, P \rangle$ is an ausp. If $X \subseteq U$ is non-empty and clopen, then $\tilde{X} = \langle X, P|X \rangle$.

CLAIM 1. Suppose that $0 \neq U' = U''$ and either P is the empty sequence or P partitiones 1. For every n, ξ there is an $n-\xi$ -uniform subausp of \tilde{U} .

PROOF: Let M, N range over non-sandy subausps of \tilde{U} and $S(M) = \{\text{Th}_\xi^n(N): N \subseteq M\}$. Choose a minimal $S = S(M_0)$ and form $t = \sum \{sw: s \in S\}$. Check that $s_1 + s_2 \in S$ if $s_1, s_2 \in S$. In virtue of Claim 4 in § 6 that implies $t \in S$. For any $s \in S$ there is $s' \in S$ such that $s = s' + t = t$. Hence M_0 is $n-\xi$ -uniform.

A congruence E on a 0-uniform ausp \tilde{U} will be called *normal* if every sandy fiber of E is singleton and for each non-empty member P_i of P the derivative of $\{x \in P_i: x/E \text{ is singleton}\}$ is equal to $\cup (U/E)'$.

CLAIM 2. For each family A of open sets in a 0-uniform ausp \tilde{U} there is a normal congruence on \tilde{U} crumpling A .

PROOF: Crumble A by an arbitrary congruence E , split every sandy non-singleton fiber of E into new singleton fibers, and for each fiber F of E and each non-empty P_i pinch out a new singleton fiber from $F \cap P_i$.

We say that $t \in \text{Tr}_\xi^n(l)$ is *sandy* if the 0-trace of t implies $l \leq 0$. The recursive function $\eta = T(n, \xi, l)$ from Theorem 3 in § 6 is used below.

LEMMA 3. Let K be a class of ausps of weight l closed under subausps. A set $S \subseteq \text{Tr}_\xi^n(l)$ includes $\text{Th}_\xi^n(K) = \{\text{Th}_\xi^n(M): M \in K\}$ if

- (0) S contains the $n-\xi$ -theory of any singleton member of K ,
- (1) S is closed under discrete sums,
- (2) $t(\omega+1) \in S$ if $t \in S$, and
- (3) $\text{Th}_\xi^n(\tilde{U}) \in S$ if $\tilde{U} \in K$ and there is a normal congruence E on \tilde{U} s.t.

$\text{Th}_\xi^n(\tilde{X}) \in S$ for any open fiber of E and $[\tilde{U}/E]_\xi^n$ is $n-T(n, \xi, l)$ -uniform.

If $U' = U''$ for every $\tilde{U} \in K$, then (2) can be replaced by a weaker condition.

- (2') $t(\omega+1) \in S$ if $t \in S$ and t is not sandy.

PROOF: Let $\tilde{U} \in K$, and X, Y range over clopen subspaces of U , and $\text{th}(X) = \text{Th}_\xi^n(\tilde{X})$. Call X *good* if $\text{th}(Y) \in S$ for each $Y \subseteq X$. By contradiction suppose that U is not good.

By Claim 2 the family of good sets can be crumbled by a normal congruence E . Let $I = U/E$. Use (0) to check that each open $i \in I$ is good. If I is discrete then each X is the discrete sum of non-empty $i \cap X$; hence, by (1), S contains each $\text{th}(X)$, hence U is good. Therefore $I' \neq 0$.

If $i \in I' - I''$, choose a clopen nbd J of i whose derivative is singleton. Given $X \subseteq \bigcup J$, use (1) and (2) to check that $\text{th}(X) \in S$. Hence $\bigcup J$ is good. But then i should be isolated in I . Therefore $I' = I''$.

By Claim 1 there is a clopen $J \subseteq I$ such that the corresponding subausp of $[\tilde{U}/E]_\xi^n$ is $n-T(n, \xi, l)$ -uniform. By (3) S contains $\text{th}(\bigcup J)$. It is easy to see that $\text{th}(X) \in S$ for any $X \subseteq \bigcup J$, i.e. $\bigcup J$ is good. But then J should be sandy which is impossible. \square

We define now uniform $n-\xi$ -theories of an ausp \tilde{U} of weight l . $UT_\xi^0(\tilde{U}) = \text{Th}^0(\tilde{U})$ and $UT_\xi^{n+1}(\tilde{U}) = \{UT_\xi^n(M) : M \in Q_\xi^{n+1}(\tilde{U})\}$ where $\eta = T(n, \xi, l)$ and $Q_\xi^{n+1}(\tilde{U})$ is the collection of $n-\eta$ -uniform ausps $\langle V/E, R \rangle$ where V is a clopen subspace of U , E is a normal congruence of V , $R = \langle Rt : t \in \text{Tr}_\xi^n(l+\xi n) \rangle$ and for every $X \in Rt$, $\text{Th}^0(\tilde{X})$ is the l -trace of t .

THEOREM 4. *There is an algorithm computing $\text{Th}_\xi^n(\tilde{U})$ from n, ξ, l and $UT_\xi^n(\tilde{U})$ whenever \tilde{U} is an $n-\xi$ -uniform ausp of weight l .*

PROOF: The algorithm uses a recursion in n . Let $\tilde{U} = \langle U, P \rangle$ be an $(n+1)-\xi$ -uniform ausp of weight l . Given $n+1, \xi, l$ and $UT_\xi^{n+1}(\tilde{U})$, compute $\eta = T(n, \xi, l)$ and $B = \{\text{Th}_\xi^n(M) : M \in Q_\xi^{n+1}(\tilde{U})\}$. Let K be the set of ausps $\langle \tilde{V}, Q \rangle = \langle V, (P|V)^\wedge Q \rangle$ where V is a clopen subspace of U and $lh(Q) = \xi n$. It suffices to compute $\text{Th}_\xi^n(K) = \{\text{Th}_\xi^n(M) : M \in K\}$, because $\text{Th}_\xi^{n+1}(\tilde{U}) = \{t \in \text{Th}_\xi^n(K) : t \text{ is not sandy}\}$. The set S_0 of $n-\xi$ -theories of singleton members of K is computable from n, ξ, l and B .

If $S = \text{Th}_\xi^n(K)$, then

(a) $S_0 \subseteq S$, S is closed under discrete sums, and $t(\omega+1) \in S$ for every non-sandy $t \in S$, and

(b) For every $\langle V/E, R \rangle \in Q_\xi^{n+1}(\tilde{U})$ with $\{t : Rt \neq 0\} \subseteq S$ there is Q such that $R = (\langle \tilde{V}, Q \rangle / E)_\xi^n$.

Clause (a) is clear. To check (b), for every $X \in Rt$ choose Q_X such that $\text{Th}_\xi^n(\tilde{X}, Q_X) = t$. The desired Q satisfies $Q|X = Q_X$ for each open fiber X of E .

Given $b = \text{Th}_\xi^n(\langle V/E, R \rangle) \in B$, we can compute $\{t : Rt \neq 0\}$, the spectrum of b . If $R = (\langle \tilde{V}, Q \rangle / E)_\xi^n$ then the algorithm of Theorem 3 in § 6 computes $\text{Th}_\xi^n(V, Q)$ from $n, \xi, l+\xi n$ and b . Therefore (b) implies

(c) If $b \in B$ and S includes the spectrum of b , then the algorithm of Theorem 3 in § 6 is applied to $\langle n, \xi, l+\xi n, b \rangle$ and the result belongs to S .

By Lemma 3, $\text{Th}_\xi^n(K)$ is the least subset of $\text{Tr}_\xi^n(l+\xi n)$ satisfying (a), (c). Hence it is computable from n, ξ, l and B .

8. Elimination of quantifiers

Let $\tilde{U} = \langle U, P \rangle$ be a 0-uniform ausp and let $\tilde{V} = \langle V, P|V \rangle$ be a subausp of \tilde{U} . It is easy to see that if $\langle V/E, R \rangle \in Q_0^1(\tilde{U})$ (where 0 denotes the zero sequence), then $R = (\tilde{V}/E)^0$ and $\langle V/E, R \rangle = [\tilde{V}/E]^0$.

CLAIM 1. $UT_0^1(\tilde{U})$ is computable from $\text{Th}^0(\tilde{U})$.

PROOF: The identity relation id on U gives $[\tilde{U}/\text{id}]^0 \in Q_0^1(\tilde{U})$. By Claim 4 in § 3 there is $C \subseteq U'$ perfect and nwd in U' . By Claim 2 in § 7 $U - C$ can be crumbled by a congruence e normal on \tilde{U} . $[\tilde{U}/e]^0$ is another member of $Q_0^1(\tilde{U})$. Let $[\tilde{V}/E]^0 \in Q_0^1(\tilde{U})$. If E is the identity relation on V , then $\text{Th}^0[\tilde{V}/E]^0 = \text{Th}^0[\tilde{U}/\text{id}]^0$. Otherwise $\text{Th}^0[\tilde{V}/E]^0 = \text{Th}^0[\tilde{U}/e]$. So $UT_0^1(\tilde{U})$ comprises exactly two elements both computable from $\text{Th}^0(\tilde{U})$.

CLAIM 2. $UT_\xi^n(\tilde{U})$ is computable from n , ξ and $\text{Th}^0(\tilde{U})$.

PROOF: The algorithm uses a recursion in n . Given $n+1$, ξ and $\text{Th}^0(\tilde{U})$ compute the weight l of U and $\eta = T(n, \xi, l)$. For $M = [\tilde{V}/E]^0 \in Q_0^1(\tilde{U})$ let $KM = \{\langle W/e, R \rangle \in Q_\xi^{n+1}(\tilde{U}): W \text{ is a clopen subspace of } V \text{ and } e = E/W\}$ and $UT_\eta^n(KM) = \{UT_\eta^n(N): N \in KM\}$. It suffices to prove that $UT_\eta^n(KM)$ is computable from $\text{Th}^0(M)$, because

$$UT_\xi^{n+1}(\tilde{U}) = \bigcup \{UT_\eta^n(KM): M \in Q_0^1(\tilde{U})\}$$

and, by Claim 1, $\{\text{Th}^0(M): M \in Q_0^1(\tilde{U})\}$ is computable from $\text{Th}^0(\tilde{U})$.

Given $M = [\tilde{V}/E]^0 \in Q_0^1(\tilde{U})$, compute $S = \{\text{Th}^0(X, P|X): X \text{ is an open fiber of } E\}$. Let $s^* = \{t \in \text{Tr}_\xi^n(l+\xi n): s \text{ is the } l\text{-trace of } t\}$ and F be the set of functions f with $S = \text{dom}(f)$ and $0 \neq fs \subseteq s^*$ for each $s \in S$. By the induction hypothesis the uniform $n-\eta$ -theory of $N = \langle W/e, R \rangle \in KM$ is computable from $\text{Th}^0(N)$ which is computable from the function f_N associating $\{t \in s^*: Rt \neq 0\}$ with each $s \in S$. It suffices to prove that for each $f \in F$ there is $N \in KM$ with $f_N = f$. Using (Cr2), split each non-empty $(\tilde{V}/E)^0$'s into disjoint subsets Qt with the same derivative where $t \in fs$. By Claim 1 in § 7, there is an $n-\eta$ -uniform subausp N of the ausp $\langle \tilde{V}/E, Q \rangle$. Clearly, $N \in KM$ and $f_N = f$.

CLAIM 3. \tilde{U} is $n-\xi$ -uniform.

PROOF: By Claim 1 in § 7 there is an $n-\xi$ -uniform subausp \tilde{V} of \tilde{U} , by Claim 2 and Theorem 4 in § 7 all $n-\xi$ -uniform subausps of \tilde{U} have the same $n-\xi$ -theory u . Let X, Y range over clopen subspaces of U , let K be the collection of subausps of \tilde{U} , $\text{th}(X) = \text{Th}_\xi^n(\tilde{X})$, $S_1 = \{\text{th}(X): X \text{ is singleton}\}$, S_2 is the closure of S_1 under discrete sums and $S = S_2 \cup \{u\}$. It suffices to verify conditions (0), (1), (2') and (3) of Lemma 3 in § 7.

(0) is trivial. In order to verify (1) check that $u+u = u$ and $u+s = u$ for each $s \in S_1$. The latter implies $u+s = u$ for each $s \in S_2$.

Checking (2'). Pick $x \in V'$ and crumble $V - \{x\}$ by a normal congruence of \tilde{V} . Together with the previous paragraph it gives $u(\omega + 1) = u$.

Checking (3). Let E be a normal congruence on \tilde{X} such that $\text{th}(Y) \in S$ for each open fiber Y of E and $[\tilde{X}/E]_\xi^n$ is $n-\eta$ -uniform where $\eta = T(n, \xi, l)$ and l is the weight of \tilde{U} . By Claim 2 \tilde{U} and \tilde{V} have the same $(n+1)-\xi$ -uniform theory; hence there is $\langle Y/e, R \rangle \in Q_\xi^{n+1}(V)$ whose $n-\eta$ -uniform theory coincides with that of $[\tilde{X}/E]_\xi^n$. It is easy to see that $\langle Y/e, R \rangle = [\tilde{Y}/e]_\xi^n$. By Theorem 3 in § 6 $\text{th}(X) = t(Y) = u \in S$.

THEOREM 4. $\text{Th}_\xi^n(\tilde{U})$ is computable from n , ξ and $\text{Th}^0(\tilde{U})$.

PROOF: Use Claims 2, 3 and Theorem 4 in § 7.

COROLLARY 5.

- (i) *The adjusted first-order theory of crumbly spaces is decidable.*
- (ii) *The first-order theory of crumbly spaces is decidable.*
- (iii) *Every two 0-uniform crumbly spaces have the same adjusted first-order theory.*
- (iv) *Every two crumbly spaces without isolated points have the same first-order theory.*

PROOF: (i) and (iii) follow from Theorem 4. Now use Corollary 4 in § 5.

References

- GUREVICH, Y., 1979, *Modest theory of short chains I*, Journal of Symbolic Logic, vol. 44, pp. 481–490
 HENSON, C. W., C. G. JOCKUSCH, JR., L. A. RUBEL, G. TAKEUTI, 1977, *First order topology*, Dissertationes Mathematicae, vol. 143

RESIDUE FIELDS OF MODELS OF P

ANGUS MACINTYRE *

University of Aberdeen, Aberdeen, U. K.

0. Introduction

This paper is intended as a contribution to diophantine model theory. I use Ax's model-theoretic analysis of finite fields to study residue fields of arbitrary models of first-order Peano arithmetic P . From 'pure' model theory I use recursive saturation. The deepest results are:

THEOREM. *If $M \models P$ and I is a nonstandard maximal ideal in M , M/I is not a recursive field.*

This is a major strengthening of Tennenbaum's Theorem.

THEOREM. *If $M \models P$, α infinite prime, a infinite, $\lambda > 1$, then in $[a, \lambda a]$ there is a prime β with $M/\alpha \equiv_{\infty, \omega} M/\beta$.*

This is a 'Bertrand Postulate' for types of primes.

Notation. M is a model of P , D the domain generated by M , L the quotient field. If necessary, M_1 , D_1 , L_1 may also be used. P is the set of irreducible elements of M . P_0 is the set of standard irreducible elements.

* Partially supported by N. S. F. The paper was written in Warsaw, while the author enjoyed the hospitality of Warsaw University and the Polish Academy of Sciences. Special thanks are due to Cecylia Rauszer and Leszek Pacholski for arranging my stay. The presentation of the paper has been improved via suggestions of Boffa, Cherlin, Van den Dries, Mc Aloon and Poizat.

1. Elementary properties of residue rings

1.1. The results of this section depend on nothing more complex than the Euclidean algorithm, Chinese Remainder Theorem, and the basic theory of recursive saturation. The best references are CHERLIN (1975), LESSAN (1978), and SCHLIFF (1978).

First, folklore:

LEMMA 1.

- (a) *Any irreducible element of M is prime;*
- (b) *Any definable ideal of D is principal;*
- (c) *Any principal prime ideal $\neq 0$ of D is maximal.*

PROOF: Exercise. \square

If $\alpha \in D$, one can of course interpret $D/\langle\alpha\rangle$ in D . A crucial special feature is:

LEMMA 2. *$D/\langle\alpha\rangle$ is finite in the sense of M , if $\alpha \neq 0$.*

PROOF: Interpret $D/\langle\alpha\rangle$ as remainders modulo α . \square

So $D/\langle\alpha\rangle$ is small in the sense of the universe M (construed as a universe of sets). Whence one has a first-order definition of satisfaction for $D/\langle\alpha\rangle$, and the next lemma follows from the general nonsense of recursive saturation (SCHLIFF, 1978):

LEMMA 3. *If M is nonstandard, and $\alpha \neq 0$, then $D/\langle\alpha\rangle$ is recursively saturated.*

Similarly,

LEMMA 4. *Suppose M nonstandard, α and β nonzero. If $D/\langle\alpha\rangle \equiv D/\langle\beta\rangle$, then $D/\langle\alpha\rangle \equiv_{\infty,\omega} D/\langle\beta\rangle$.*

In neither lemma is the hypothesis that M is nonstandard necessary. Another folklore remark is:

LEMMA 5.

- (a) *If $\alpha \neq 0$ and α standard, then $D/\langle\alpha\rangle$ is canonically isomorphic to $\mathbb{Z}/\langle\alpha\rangle$;*
- (b) *If α is nonstandard, then $D/\langle\alpha\rangle$ is a ring of characteristic 0.*

PROOF: Trivial. \square

1.2. Mostly I look at nonstandard prime α , in which case $D/\langle\alpha\rangle$ is a field of characteristic 0. To understand the structure of those fields

requires much deeper number theory, namely the Lang-Weil estimates (LANG, WEIL, 1954) and Čebotarev Theorem (LANG, 1970). However, the following basic property is easily proved:

LEMMA 6. *If α is prime, $D/\langle\alpha\rangle$ has exactly one extension of degree n , for each integer $n \geq 1$.*

PROOF: Formalize and prove in P the argument in HERSTEIN (1964). \square

Note, in connection with Lemma 6, that a stronger conclusion holds. For the version of Lemma 6 in which one quantifies over all $n \in M$ is provable in P , I did not investigate the significance of extensions of non-standard degree.

1.3. Cherlin's correspondence. It has been known for a long time that the ideal theory of a general D lacks most of the structure of the classical case $D = \mathbf{Z}$. For example, no nonstandard D is noetherian, and not every nonzero prime ideal is maximal. Nevertheless, some very useful structure theory is known. My source is CHERLIN (1975). Unfortunately he does not consider general models of P , but since he uses only the Chinese Remainder Theorem, it is evident that all his results generalize to the present setting. (I remark that all my results easily generalize to the obvious theories of algebraic integers.)

I now review his results. If $a \in M$, let $S_a = \{p \in P: p \mid a\}$. Then S_a is definable, and is finite in the sense of M for $a \neq 0$. Conversely, if X is a subset of P , definable in M , and finite in the sense of M , then $X = S_a$ for some a . (This argument fails in models of bounded induction. See PARIKH, 1971, for the explanation.) Let $\text{Def}(P)$ be the Boolean algebra of all subsets of P definable in M . Then the S_a are precisely those elements of $\text{Def}(P)$ finite in the sense of M . A filter μ on $\text{Def}(P)$ is *bounded* if it contains an S_a with $a \neq 0$.

Suppose I is a nonzero ideal in D . Let $\mu(I) =$ the filter generated by all S_a , $a \in I$. Then $\mu(I)$ is a bounded filter. Conversely, let μ be a bounded filter on $\text{Def}(P)$. Let $I(\mu) = \{a: S_a \in \mu\}$. Then $I(\mu)$ is a nonzero ideal in D . The basic facts are:

LEMMA 7.

- (a) I prime $\Rightarrow \mu(I)$ is an ultrafilter;
- (b) μ a bounded ultrafilter $\Rightarrow I(\mu)$ maximal;
- (c) I maximal $\Rightarrow I(\mu(I)) = I$;
- (d) μ a bounded ultrafilter $\Rightarrow \mu(I(\mu)) = \mu$;
- (e) $I \neq 0$ prime $\Rightarrow I(\mu(I))$ is the unique maximal ideal $\supseteq I$;

- (f) $I \neq 0$ principal prime $\Rightarrow \mu(I)$ principal;
- (g) μ a principal ultrafilter $\Rightarrow I(\mu)$ principal prime.

PROOF: CHERLIN (1975). \square

In addition to this representation of general maximal ideals, Cherlin obtained a more complex representation of prime ideals. I will not use this here.

Per se, the above representation is of limited value since ultrafilters are admittedly farout objects. But now in the case of models of P the Chinese Remainder Theorem gives a special twist.

Let $D^{(P)}$ be the ring of definable functions from P to D , and let $\prod_{(P)} D/\langle P \rangle$ be the ring of definable functions f on P with $f(p) \in D/\langle p \rangle$ for $p \in P$. Let μ be an ultrafilter on $\text{Def}(P)$. Then one forms respectively the “definable ultrapower” $D^{(P)}/\mu$, and the “definable ultraproduct” $(\prod_{(P)} D/\langle P \rangle)/\mu$. (See

CHERLIN, 1975.) One has the usual Łoś Theorem that the diagonal embedding $\Delta_\mu: D \rightarrow D^{(P)}/\mu$ is elementary.

Cherlin’s striking observation (using only the Chinese Remainder Theorem) is:

THEOREM 1. *Suppose that I is maximal in D . Then (the class of) id_P is prime in $D^{(P)}/\mu(I)$, $\langle \text{id}_P \rangle \cap \Delta_{\mu(I)}(D) = I$, and the induced homomorphism*

$$D/I \rightarrow (D^{(P)}/\mu(I))/\langle \text{id}_P \rangle$$

is an isomorphism.

So, I is made principal in the elementary extension $D^{(P)}/\mu(I)$. This is a powerful idea in the analysis of D/I . In particular, we may conclude:

LEMMA 8. *Suppose that I is maximal. Then D/I is recursively saturated, and has exactly one extension of each degree $<\omega$.*

PROOF: Immediate from Lemmas 3 and 6. \square

It is, however, important to note that one cannot make two ideals principal simultaneously, and we shall see that Lemma 4 does not extend to general maximal ideals.

Heuristically, it is important to represent $(D^{(P)}/\mu(I))/\langle \text{id}_P \rangle$ as an ultraproduct.

LEMMA 9. *Canonically, $(D^{(P)}/\mu(I))/\langle \text{id}_P \rangle \simeq (\prod_{(P)} D/\langle P \rangle)/\mu(I)$.*

PROOF: Exercise. \square

2. Applying Ax's results

2.1. Ax (1968) made a deep study of ultraproducts of finite fields, providing for example elementary invariants for all such fields. First, one observes some elementary properties of finite fields, namely:

- (a) they are perfect;
- (b) for each $n \geq 1$ they have exactly one extension of degree n .

These properties are of course preserved under ultraproducts. Note that if α is prime, $D/\langle\alpha\rangle$ has the above properties (Lemma 5(b), and Lemma 6).

The key property discovered by Ax is that any infinite ultraproduct of finite fields is pseudo-algebraically closed (abbreviated p.a.c.). I recall:

DEFINITION. K is p.a.c. if every absolutely irreducible variety V defined over K has a K -valued point.

DEFINITION. K is pseudofinite if K is p.a.c. and satisfies (a) and (b) above.

It is easily seen that no finite field is pseudofinite. One has, however, the following deep result:

THEOREM 2. Any infinite ultraproduct of finite fields is pseudofinite.

PROOF: See Ax (1968). \square

The proof depends on Weil's Riemann Hypothesis for curves over a finite field. That result says:

THEOREM 3. If \mathcal{C} is an absolutely irreducible curve of genus g defined over F_q , and N_m is the number of points of \mathcal{C} in F_{q^m} , then

$$|N_m - (q^m + 1)| \leq 2gq^{m/2}.$$

PROOF: See BOMBIERI (1976) or SCHMIDT (1976) for elementary proofs. \square

From this it is easily deduced that any infinite ultraproduct K of finite fields satisfies:

(*) Every absolutely irreducible curve over K has infinitely many K -valued points.

Finally, as pointed out to me by Van den Dries, one has the easy:

LEMMA 10. Any field satisfying (*) is p.a.c.

Now I want to show that if α is a nonstandard prime then $D/\langle\alpha\rangle$ is pseudofinite. So it suffices to show that $D/\langle\alpha\rangle$ satisfies (*). Now by ident-

ifying curves with defining polynomials $f(x, y) = 0$, and using the well-known result that the concept *absolutely irreducible* is quantifier-free definable, one is essentially reduced to proving Theorem 3 in P . It is clear that the elementary Riemann-Roch theory is formulable and provable in P , together with the elementary estimates linking genus and degree. Then inspection of BOMBIERI's (1976) proof shows it to be "in P ". Now to get $D/\langle\alpha\rangle$ to satisfy (*), I argue thus. Let \mathcal{C} be absolutely irreducible over $D/\langle\alpha\rangle$, defined by $f(x, y) = 0$, f of standard degree, n say. Then the genus is provable finite. So by Theorem 3

$$|N_1 - (|\alpha| + 1)| \leq 2g|\alpha|^{1/2}.$$

Since $|\alpha|$ is infinite, so is N_1 , i.e. there are infinitely many points on \mathcal{C} in $D/\langle\alpha\rangle$. I conclude:

LEMMA 11. *For nonstandard prime α , $D/\langle\alpha\rangle$ is p.a.c.*

2.2. Generalizing Tennenbaum's Theorem. Lemma 11 leads to a very striking conclusion.

THEOREM 4. *Suppose that α is a nonstandard prime, and $D/\langle\alpha\rangle$ is countable. Then $D/\langle\alpha\rangle$ is not a recursive field.*

PROOF: In MACINTYRE (forthcoming) I show that no recursively saturated pseudofinite field of characteristic 0 is recursive. Now apply Lemma 8 and 11. \square

Tennenbaum (see EHRENFEUCHT, KREISEL, 1968) showed that if M is nonstandard, then M is not recursive. (And then D is not recursive, by MACINTYRE 1949.) But heuristically appealing to a 'Hasse principle', one would have hoped that $D/\langle\alpha\rangle$ might be recursive.

I remark that my proof in MACINTYRE (forthcoming) depends on some delicate points in the work of Ax.

2.3. Evidently we may apply Cherlin's idea to extend both Lemma 11 and Theorem 4 to general maximal ideals.

2.4. I now use finer details of Ax's analysis to investigate the following problems:

- (i) *What are the constraints on K in order that $K \simeq D/\langle\alpha\rangle$ for some D, α ?*
- (ii) *Given D , what K occurs as D/I where I is maximal?*
- (iii) *Given $\alpha \in D$, how are the β distributed in D so that $D/\langle\alpha\rangle \equiv D/\langle\beta\rangle$?*

For (i) and (ii) I will not need to work in P . Rather I proceed model-theoretically. But for (iii) I will have to work in P , formalizing a somewhat

difficult argument of analytic number theory. The *sine qua non* for all my results is Čebotarev's Theorem (LANG, 1970).

2.5. *The elementary invariants of a pseudofinite field.* For convenience, fields are of characteristic 0, unless otherwise stated.

DEFINITION.

- (a) $\text{Abs}(K) = K \cap \tilde{Q}$;
- (b) $A_+(K) = \{f \in Q[x]: f \text{ solvable in } K\}$;
- (c) $A_-(K) = \{f \in Q[x]: f \text{ unsolvable in } K\}$;

LEMMA 12. $\text{Abs}(K_1) = \text{Abs}(K_2) \Leftrightarrow A_+(K_1) = A_+(K_2) \Leftrightarrow A_-(K_1) = A_-(K_2)$.

PROOF: Ax (1968). \square

THEOREM 5. Suppose that K_1 and K_2 are pseudofinite. Then $K_1 \equiv K_2 \Leftrightarrow \text{Abs}(K_1) \simeq \text{Abs}(K_2)$.

PROOF: Ax (1968). \square

With this I make a small beginning on (iii). Suppose that α is a non-standard prime in D , and $[a, b]$ is an interval in M . Then

LEMMA 13. There exists a prime $\beta \in [a, b]$ with $D/\langle \alpha \rangle \equiv D/\langle \beta \rangle \Leftrightarrow$ for all finite U, V with $U \subseteq A_+(D/\langle \alpha \rangle), V \subseteq A_-(D/\langle \alpha \rangle)$ there is a prime $\beta \in [a, b]$ with $U \subseteq A_+(D/\langle \beta \rangle), V \subseteq A_-(D/\langle \beta \rangle)$.

PROOF: By Theorem 5 and a routine “overspill” argument. \square

The remaining problem is to decide when there is a prime in $[a, b]$ satisfying the conditions given by U, V . This will be taken up in Section 3. For now I look at general sets of primes defined by conditions U, V as above.

DEFINITION. The *Ax algebra* $\text{Ax}(M)$ of M is the Boolean subalgebra of $\text{Def}(P)$ generated by all sets $\{\alpha \in P: D/\langle \alpha \rangle \models (\exists x)(f(x) = 0)\}$, where $f \in Q[x]$.

In particular, for finite $U, V \subseteq Q[x]$,

$$\{\beta \in P: U \subseteq A_+(D/\langle \beta \rangle), V \subseteq A_-(D/\langle \beta \rangle)\} \in \text{Ax}(M).$$

A basic fact about $\text{Ax}(N)$ is:

LEMMA 14. Every member of $\text{Ax}(N)$ is recursive.

PROOF: Clear. \square

The structure of elements of $\text{Ax}(N)$ is revealed by Čebotarev's Theorem. By this method, Ax answered the following question: Which subfields of $\tilde{\mathbb{Q}}$ are of the form $\text{Abs}(K)$, for K pseudofinite?

The answer is:

THEOREM 6. *Let A be a subfield of $\tilde{\mathbb{Q}}$. The following are equivalent:*

- (a) A is procyclic (i.e. has at most one extension of each degree);
- (b) $A \simeq \text{Abs}(K)$ for some pseudofinite K ;
- (c) $A \simeq \text{Abs}(K)$ for some K which is of the form $(\prod_{p \in P_0} F_p)/\mu$, where μ is

a nonprincipal ultrafilter on P_0 .

PROOF: See Ax (1968). \square

So, in particular, every pseudofinite field is elementarily equivalent to an ultraproduct of the standard finite prime fields. I now convert Ax's standard ultraproduct representation into a "Skolem ultraproduct" in nonstandard M . First note that by Theorem 5 the elementary type of $(\prod F_p)/\mu$ is uniquely determined by $\mu \cap \text{Ax}(N)$. I now have to find a way of lifting $\mu \cap \text{Ax}(N)$ to a bounded ultrafilter v on $\text{Def}(P)$ so that $(\prod_{(P)} D/\langle p \rangle)/v \equiv (\prod F_p)/\mu$. The key idea is to use Lemma 14. Since the elements of $\text{Ax}(N)$ are recursive, they have Δ_1 definitions in N , which thereby lift to M , defining sets whose intersections with N give back the original set. By this process, $\mu \cap \text{Ax}(N)$ lifts to a subset of $\text{Def}(P)$ with the finite intersection property. Indeed, if I prescribe an infinite $\alpha \in M$, the set whose members are $[0, \alpha]$ and the liftings of $\mu \cap \text{Ax}(N)$ has the finite intersection property, and so extends to a bounded ultrafilter v on $\text{Def}(P)$. By Łoś' Theorem, and Theorem 5, $(\prod_{(P)} D/\langle p \rangle)/v \equiv (\prod F_p)/\mu$. By Cherlin, there is a maximal I so that $D/I \simeq (\prod_{(P)} D/\langle p \rangle)/v$. This proves:

THEOREM 7. *For every pseudofinite field K , and every nonstandard M , there is a maximal I so that $D/I \equiv K$.*

Remarks. (a) Given D , K , I must in general be nonprincipal. For by Ax (1968) there are 2^{\aleph_0} theories of K .

(b) One may ask if I can always be chosen nonprincipal, i.e. if v can be chosen nonprincipal. This is evidently so, since each member of $\mu \cap \text{Ax}(N)$ is infinite. Thus we may conclude that for each nonstandard prime α in M there is a nonprincipal I with $D/\langle \alpha \rangle \equiv D/I$.

2.6. I now give a satisfactory, but not quite complete, answer to Problem I. Recall from SCHLIPP (1978) the notion of *resplendent* model.

THEOREM 8. Suppose that K is a resplendent pseudofinite field. Then there exists M and prime α such that $D/\langle\alpha\rangle \simeq K$.

PROOF: By Theorem 7, Cherlin, and the Keisler-Shelah Theorem (SHELAH, 1971), there exist an elementary extension K^* of K and M , α so that $D/\langle\alpha\rangle \simeq K^*$. By passing to a further elementary extension if required, I can suppose $\text{card}(M) = \text{card}(K^*)$, and indeed that $M = K^*$ qua sets. Then there exist \oplus, \odot on K^* , and $\alpha \in K^*$, so that $\langle K^*, \oplus, \odot \rangle \models P$, and there exists an isomorphism $f: K^*/\langle\alpha\rangle \simeq K^*$. Since P is recursive and K resplendent (SCHLIPP, 1978), there are $\oplus_K, \odot_K, \alpha_K, f_K$ on K with the corresponding properties. \square

COROLLARY. Suppose that K is a countable pseudofinite field. Then the following are equivalent:

- (a) K is recursively saturated;
- (b) there exists M, α with $D/\langle\alpha\rangle \simeq K$.

PROOF: For countable K , recursively saturated is equivalent to resplendent (SCHLIPP 1978). Then use Theorem 8 and Lemma 3. \square

Naturally, one wants to know if the hypothesis of countability is needed. I suppose so.

2.7. A very striking answer to Problem II would come from improving Theorem 7 by replacing \equiv by \simeq . However, no such improvement is possible, and Theorem 7 constitutes our answer to Problem II.

Let $T = \text{Th}(N)$. FEFERMAN (1958) showed that T has no arithmetical nonstandard model. This is of course a sharpening of Tennenbaum's Theorem. It turns out that there is a corresponding sharpening of my Theorem 4. In MACINTYRE (forthcoming) I prove that no Σ_n -saturated pseudofinite field has a Δ_n model.

Next, the following is, or should be, folklore:

LEMMA 15. Any structure finite in the sense of a model of T is Σ_n -saturated, for each n .

So

THEOREM 9. Suppose that $M \models T$ and α is a nonstandard prime. Then $D/\langle\alpha\rangle$ is not Δ_n , for any n .

One more piece of folklore:

LEMMA 16. Let \mathcal{T} be a theory which is Σ_n (qua set of Gödel numbers). Then \mathcal{T} has a recursively saturated model of complexity at most Δ_{n+2} .

From these pieces one constructs the following counterexample to an improvement of Theorem 7.

Example. Let \mathcal{T} be a decidable complete theory of a pseudofinite field. (There are many such; Ax, 1968.) Let K be a Δ_2 recursively saturated model of \mathcal{T} . Then there is no model M of T , and $\alpha \in M$, so that $D/\langle\alpha\rangle \simeq K$.

This is immediate from Theorem 9. We can even replace $\langle\alpha\rangle$ by a maximal I .

2.8. I now present counterexamples to any extension of Lemma 4 to the case of general maximal ideals.

THEOREM 10. *Suppose that M is countable and I is maximal in D . Then there is a maximal J such that $D/I \equiv D/J$ but $D/I \not\equiv_{\omega,\omega} D/J$.*

PROOF. Let $\mathcal{T} = \text{Th}(D/I)$. By a result in MACINTYRE (forthcoming), $S_n(\mathcal{T})$ has cardinal 2^{\aleph_0} for some n . Since D/I is countable, I can select $\tau \in S_n(\mathcal{T})$, τ not realized in D/I . The theorem will be proved by constructing J so that $D/I \equiv D/J$ and D/J realizes τ . The idea is close to that of Theorem 6. Consider the sets $\{p \in P_0 : Z/\langle p \rangle \models \exists v_1, \dots, v_n \psi\}$, for each $\psi \in \tau$. By Theorem 5, each of these sets is in $\text{Ax}(N)$. The sets have the finite intersection property. Now lift to a bounded ultrafilter on $\text{Def}(P)$, as in 2.5. \square

COROLLARY (to proof). *With the same hypotheses, there are 2^{\aleph_0} such J such that the D/J are pairwise (∞, ω) -inequivalent.*

PROOF: There are 2^{\aleph_0} choices for τ . \square

2.9. I conclude this section with a simple omitting-types result.

THEOREM 11. *Let \mathcal{T} be a complete theory of a pseudofinite field. The following are equivalent:*

(a) \mathcal{T} is decidable;

(b) For each nonstandard M there exists α such that $D/\langle\alpha\rangle \models \mathcal{T}$.

PROOF: (a) \Leftrightarrow (b). Overspill, since \mathcal{T} can be axiomatized by sentences of complexity Σ_3 .

(b) \Leftrightarrow (a). Routine omitting types. \square

3. Bertrand postulate for types

I come now to the most explicitly arithmetical result of the paper.

THEOREM 12. *Let α be a nonstandard prime. Let $a \in M$, a nonstandard. If $b \in M$, and $b > \lambda \cdot a$, where $\lambda \in Q$, $\lambda > 1$, then there is a prime β with $\beta \in [a, b]$, and $D/\langle\alpha\rangle \equiv D/\langle\beta\rangle$.*

PROOF: Lemma 13 tells us what we have to prove. We are once again in the position of having to prove something difficult "in P ". This time it is an effective version of Čebotarev's Theorem. The source proof is LAGARIAS, ODYLYZKO (1977). I now give a very brief outline of the program. To understand it one must understand Ax's decision-method for one variable sentences, at the end of Ax (1968).

Ax gives a uniformly effective method which, beginning with U , V as in Lemma 13, constructs a finite normal extension $Q^{(U,V)}$ of Q . The recursive nature of his method means that it can be formulated for a general M , providing for finite (in the sense of M) U , $V \subseteq K[x]$, a finite (in a sense of M) normal (in sense of M) extension $L^{(U,V)}$ of L . (Recall that L is the quotient field of D .)

The point of $Q^{(U,V)}$ is the following. Except for the (finite, computable) primes of Z ramified in $Q^{(U,V)}$, the property

$$U \subseteq A_+(Z/\langle p \rangle), \quad V \subseteq A_-(Z/\langle p \rangle) \quad (p \in P_0)$$

depends only on the Frobenius class of p in $G(Q^{(U,V)}|Q)$. This generalizes to $L^{(U,V)}$, replacing where appropriate *finite* by *finite in M* . One is of course obliged to do nonstandard Galois theory.

I do not claim that it is trivial that the elementary theory (pre-Čebotarev) of the Frobenius can be formalized in P . I do claim that if one understands the development in LANG's (1970), then it is evident how to proceed. Probably the use of Takeuti's conservative extension results (TAKEUTI, 1976) can shorten the formalization. One should observe that the p -adics intervene in the theory.

It is very important to note that if U , V are standard finite subsets of $Q[x]$, then $L^{(U,V)}$ has standard finite dimension over L , and $G(L^{(U,V)}|L)$ is a standard finite group. Moreover, $L^{(U,V)}$ is the splitting field of some $f \in Q[x]$, and only finitely many p , all standard, ramify.

So our task now is thus. Given $[a, b]$ and a conjugacy class \mathcal{C} in $G(L^{(U,V)}|L)$, to find a prime p in $[a, b]$ with \mathcal{C} as the Frobenius class of p . When $L = Q$, this is a problem of effective estimates in Čebotarev's Theorem, and is not answered by classical treatments. Fortunately, there now exists a clearly arranged proof, by LAGARIAS and ODYLYZKO (1977), of a good effective estimate in the general Čebotarev Theorem. Long as their proof is, it is evidently available in Takeuti's system (TAKEUTI, 1976), and so in P . For our purposes, the following version suffices:

There is a definable $f: M \times L \rightarrow M$ such that

- (i) if $x \in N$, $y \in Q$, $y > 1$, then $f(x, y) \in N$;

(ii) if L_1 is a Galois extension of L of M -finite dimension n , and if \mathcal{C} is a conjugacy class in $G(L_1|L)$, then there is a prime p in $[a, \lambda a]$ of Frobenius \mathcal{C} , provided $\lambda > 1$ and $a > f(n, \lambda)$.

If $L_1 = L^{(U, V)}$, U, V finite $\subseteq Q[x]$, then $n \in N$. So if $\lambda \in Q$, then $f(n, \lambda) \in N$, and p can be found in $[a, \lambda a]$ if a is infinite and $\lambda > 1$. This proves the theorem. \square

4. Behavior of primes under extensions

4.1. Consider an extension $M \rightarrow M_1$ of models of P . Recall that by MATIJASEVIČ (1970) the notion *prime* is preserved under such extensions. Suppose $\alpha \in M$, α prime. Let $\langle \alpha \rangle_{M_1}$ be the ideal generated by α in D_1 . The Euclidean algorithm shows that $\langle \alpha \rangle_{M_1} \cap D = \langle \alpha \rangle$, so I drop the subscript " M_1 ". There is an induced injection $D/\langle \alpha \rangle \rightarrow D_1\langle \alpha \rangle$. Now, it is known that not every complete theory of a pseudofinite field is model-complete (Ax, 1968). However:

THEOREM 13. $D/\langle \alpha \rangle \rightarrow D_1/\langle \alpha \rangle$ is elementary.

PROOF: By Ax (1968), it is enough to show that the map is "relatively algebraically closed". If $a_0, \dots, a_n \in D$, the condition

$$\neg(\exists x)(a_0 + a_1 x + \dots + a_n x^n \equiv 0 \text{ mod } \alpha)$$

as, by Matejasevič, Σ_1 , and so preserved upwards. \square

4.2. End extensions.

THEOREM 14. If $M \rightarrow M_1$ is an end extension, then

$$\{\text{Th}(D/\langle \alpha \rangle): \alpha \in M, \alpha \text{ prime}\} = \{\text{Th}(D_1/\langle \beta \rangle): \beta \in M_1, \beta \text{ prime}\}.$$

PROOF: If $M \rightarrow M_1$ is an end extension, M and M_1 have the same standard system (FRIEDMAN). Now, if one uses the fact that each complete theory \mathcal{T} of a pseudofinite field is Σ_3 -axiomatizable, the following obvious lemma completes the proof. \square

LEMMA 17. $(\exists \alpha \in M)(D/\langle \alpha \rangle \models \mathcal{T}) \Leftrightarrow \mathcal{T} \text{ is in the standard system of } M$.

5. Superspill

The following principle is for prime quotients a converse to the familiar overspill principle.

THEOREM 15. If α is prime, and $D/\langle\alpha\rangle \models \Phi$, then for some standard p $D/\langle p\rangle \models \Phi$.

PROOF: $D/\langle\alpha\rangle$ is pseudofinite, and then use Theorem 6. \square

Added in proof. After the paper was written, Leonard Lipshitz told me that Tennenbaum had outlined a proof of a special case of Theorem 4 in the early 1970's. This proof, for the case of complete arithmetic and principal ideals, was never published.

To compensate, I announce an easy refinement of the first theorem of Section 0. The $+$ of M/I can always be chosen recursive, since $M/I \cong Q^{(\omega)}$ as additive group. The \bullet of D/I can sometimes be chosen recursive. In this respect the situation differs from that of models of Peano arithmetic, where neither $+$ nor \bullet can be chosen recursive.

References

- Ax, J., 1968, *The elementary theory of finite fields*, Annals of Mathematics (2), vol. 88, pp. 239–271
- BOMBIERI, E., 1976, *Hilbert's 8th problem: An analogue*, pp. 269–274 in Proceedings of Symposia in Pure Mathematics, vol. 28 (Mathematical developments arising from Hilbert problems), (American Mathematical Society, Providence)
- CHERLIN, G., 1975; *Ideals of integers in nonstandard number fields*, in: Model Theory and Algebra, eds. D. Saracino and V. Weisspfennig (Springer Lecture Notes 498, Berlin), pp. 60–90
- EHRENFEUCHT, A., and G. KREISEL, 1966, *Strong models of arithmetic*, Bulletin de l'Academie Polonaise des Sciences, série des sciences mathématiques, astronomiques et physiques, vol. 14, pp. 107–110
- FEFERMAN, S., 1958, *Arithmetically definable models of formalized arithmetic*, Abstract 550–21, Notices of the American Mathematical Society, vol. 5, pp. 679–680
- FRIEDMAN, H., 1973, *Countable models of set theories*, in: Cambridge Summer School of Mathematical Logic, ed. Mathias, Springer Lecture Notes 337, pp. 539–573
- HERSTEIN, I., 1964, *Topics in algebra* (Bleisell, New York)
- LAGARIAS, J., and A. ODYLYZKO, 1977, *Effective versions of the Chebotarev density theorem*, in: Algebraic number fields, ed. A. Frohlich (Academic Press, London)
- LANG, S., 1970, *Algebraic number theory* (Addison-Wesley, Reading-London)
- LANG, S., and A. WEIL, 1954, *Number of points of varieties in finite fields*, American Journal of Mathematics, vol. 76, pp. 819–827
- LESSAN, H., 1978, Ph. D. Thesis (Manchester)
- MACINTYRE, A., *Some observations about types in fields*, to appear in Proceedings of University of Connecticut Meeting, ed. M. Lerman
- MATIJASEVIČ, Yu. V., 1970, *Enumerable sets are diophantine*. (Russian), Doklady Akademii Nauk SSSR, vol. 191, pp. 279–282. Engl. transl., Soviet Mathematics Doklady, vol. 11, pp. 354–358
- PARikh, R., 1971, *Existence and feasibility in arithmetic*, The Journal of Symbolic Logic, vol. 36, pp. 494–508

- ROBINSON, J., 1949, *Definability and decision problems in arithmetic*, The Journal of Symbolic Logic, vol. 14, pp. 98–114
- SCHLIFF, J., 1978, *Towards model theory through recursive saturation*, The Journal of Symbolic Logic, vol. 43, pp. 183–206
- SCHMIDT, W., 1976, *Equations over finite fields*, Springer Lecture Notes 536 (Berlin)
- SHELAH, S., 1971, *Every two elementarily equivalent models have isomorphic ultrapowers*, Israel Journal of Mathematics, vol. 10, pp. 224–233
- TAKEUTI, G., 1976, *A conservative extension of Peano arithmetic*, preprint (Urbana)

NUMBER OF MODELS IN COMPLETE VARIETIES

E. A. PALYUTIN

Institute of Mathematics, Novosibirsk, U.S.S.R.

1. Introduction

A class M of algebraic systems of a signature Σ , closed with respect to subsystems, Cartesian products and homomorphic images is called a *variety*. A variety M is called *complete* if there exist infinite M -systems and all of them are elementary equivalent. All the varieties considered below have signature of finite or countable power.

If K is a class of algebraic systems, then by $S_K(\kappa)$ we denote the number of isomorphic types of K -systems of power κ . Such a correspondence S_K we call the *spectrum* of K .

The aim of the present paper is to describe all the spectra S_M for complete varieties M . This description is obtained from the following theorem.

THEOREM 1. *Spectrum of a complete variety has one of the following values S_0 , S_1 , $S_{m,p}$, where $m \in \omega$; p is a prime number:*

$$S_{m,p}(\kappa) = \begin{cases} 1 & \text{if } \kappa = 1 \text{ or } \kappa \geq \omega, \\ 1 & \text{if } \kappa = p^{m+\tau} \text{ for some } \tau \in \omega, \\ 0 & \text{otherwise;} \end{cases}$$
$$S_1(\kappa) = \begin{cases} 1 & \text{if } \kappa = 1 \text{ or } \kappa > \omega, \\ \omega & \text{if } \kappa = \omega, \\ 0 & \text{otherwise;} \end{cases}$$
$$S_0(\kappa) = \begin{cases} 1 & \text{if } \kappa = 1, \\ 2^\kappa & \text{if } \kappa \geq \omega, \\ 0 & \text{otherwise.} \end{cases}$$

For definitions of all the notions used here see CHANG and KEISLER (1973) or ERSHOV and PALYUTIN (1979). Power of a set X we denote by $|X|$. Collection $\langle w_1, \dots, w_n \rangle$ we denote by \bar{w} , and we just write $\bar{w} \in X$ instead of $w_1 \in X, \dots, w_n \in X$. If K is a class of algebraic systems, then by K_+ and K_∞ we denote a class of not one-element K -systems and the class of infinite K -systems, respectively.

A class of algebraic systems K is called *categorical in power κ* if all the K -systems of power κ are isomorphic. In particular, if there are no K -systems of power κ , then, by definition, K is categorical in κ .

Algebraic systems will be denoted below by the Gothic letters \mathfrak{A} and \mathfrak{B} , and their carriers by the corresponding Latin letters A and B . Formula Φ which is equivalent to the formula of $\exists x_1 \dots \exists x_n \Psi$ kind, where Ψ is the conjunction of atomic formulae, is called \exists^+ -formula. For simplicity of denotations we identify with \mathfrak{A} a subsystem of Cartesian power \mathfrak{A}^I which consists of functions f_a identically equal to a , $a \in A$. In particular, f_a we denote by a . If $\Phi(x, \bar{y})$ is \exists^+ -formula, \mathfrak{A} is an algebraic system and $\bar{a} \in A$, then the denotation $\Phi(x, \bar{a})$ is also called \exists^+ -formula. If $\Phi(x, \bar{y})$ is a formula, \mathfrak{A} an algebraic system, and $\bar{a} \in A$, then by $\Phi(\mathfrak{A}, \bar{a})$ we denote the set

$$\{b \in A \mid \mathfrak{A} \models \Phi(b, \bar{a})\}.$$

If $\Phi(x, y)$ is \exists^+ -formula, then the set $\Phi(\mathfrak{A}, \bar{a})$ is called \exists^+ -set (in \mathfrak{A}). If the proposition Φ is true on all the systems from the class K , then we write $K \models \Phi$.

The set $X \subseteq A$ is called *strongly \exists^+ -minimal (in the system \mathfrak{A})* if for any \exists^+ -formula $\Phi(x, \bar{y})$ of system's signature and for any $\bar{a} \in A$ the set $X \cap \Phi(\mathfrak{A}, \bar{a})$ is empty, one-element or coincides with X .

Remark. If M is a complete variety of signature Σ , then for any predicate symbol P of arity n we have $M \models \forall x_1 \dots \forall x_n P(x_1, \dots, x_n)$. Indeed, let \mathfrak{A} be an M_∞ -system and let \mathfrak{A}' be obtained from \mathfrak{A} by substituting the value of the predicate P for the identically true one. Then \mathfrak{A}' is the homomorphic image of \mathfrak{A} , and consequently, \mathfrak{A}' belongs to M_∞ . From the completeness of M it follows that $M_\infty \models \forall x_1 \dots \forall x_n P(x_1, \dots, x_n)$. If on some M -system \mathfrak{B} the predicate P has not been identically true, then P would not be identically true on M_∞ -system $\mathfrak{A} \times \mathfrak{B}$, which is impossible.

From the above remark it follows that for any complete variety M we have $S_M(1) = 1$.

The proof of Theorem 1 will be held in the following three sections.

2. Locally finite varieties

A. LACHLAN (1972) proved that if a complete variety M has a finite not one-element system, then M is countable categorical. E. PALYUTIN proved (ABAKUMOV *et al.*, 1972) that a countable categorical variety is categorical in all powers. Thus, in order to fully describe the spectra of complete varieties M , having finite not one-element systems, it is sufficient to describe for such M the sets

$$W(M) = \{n \in \omega \mid \text{exists } M\text{-system of power } n\}$$

For quasivarieties such sets have been described in PALYUTIN (1973) and PALYUTIN (1975) with full proof. From the results of PALYUTIN (1975) it easily follows that for a complete variety M there exist a number $m \in \omega$ and a prime number p such that

$$W(M) = \{p^{mr} \mid r \in \omega\}.$$

Note that due to the Ryll-Nardzewski theorem (cf. CHANG and KEISLER, 1973) the countable categorical variety has a finite M_+ -system. Thus, the following theorem holds:

THEOREM 2.1. *For a complete variety M the following conditions are equivalent:*

- (1) M contains a finite, not one-element system,
- (2) $S_M(\omega) = 1$,
- (3) S_M coincides with $S_{m, p}$ for some $m \in \omega$ and prime p .

3. Uncountable spectrum

Let T be a complete countable elementary theory of signature Σ , let \mathfrak{U} be a model of T , and $X \subseteq A$. By \mathfrak{U}_X we denote the enrichment of \mathfrak{U} up to the system of signature containing constants c_a , $a \in X$ and the value of c_a in \mathfrak{U}_X is a . The theory T is called *stable in power κ* if for any of its models \mathfrak{U} and for any set $X \subseteq A$ from $|X| \leq \kappa$ it follows that the power of 1-types, consistent with the theory $\text{Th}(A_X)$ does not exceed κ . The theory T is called *stable* if it is stable in some infinite power. The theory T is called *super-stable* if it is stable in any power $\geq 2^\omega$.

We shall make use of the following well-known theorems:

MORLEY'S THEOREM (cf. CHANG and KEISLER, 1973). *If an axiomatizable class K is categorical in some uncountable power, then K is categorical in all uncountable powers.*

SHELAH'S THEOREM (cf. SHELAH, 1978). *If a complete countable theory T is not superstable, then for any $\kappa > \omega$ T has 2^κ isomorphism types of models of power κ .*

For us the theorem proved by the author (PALYUTIN, 1979) is also important.

THEOREM 3.1 (PALYUTIN, 1979). *For a complete variety M the following conditions are equivalent:*

- (1) *M is categorical in uncountable powers;*
- (2) *in any M -system \mathfrak{U} there does not exist an infinite strictly decreasing sequence of \exists^+ -sets.*

Due to Morley's and Shelah's theorems, to describe spectra of complete varieties in uncountable powers, it is sufficient to prove the following theorem.

THEOREM 3.2. *For a complete variety M the following conditions are equivalent:*

- (1) *M is categorical in uncountable powers;*
- (2) *$\text{Th}(M_\infty)$ is superstable.*

We first prove several lemmas. In the rest of the section, by M we denote a complete variety of signature Σ , all the systems under consideration are M -systems, and the formulae have the signature Σ .

LEMMA 3.3 (PALYUTIN, 1979). *If $\text{Th}(M_\infty)$ is stable, then for any \exists^+ -formula $\Phi(x, \bar{y})$, system \mathfrak{U} and $\bar{a}, \bar{b} \in A$, the sets $\Phi(\mathfrak{U}, \bar{a})$ and $\Phi(\mathfrak{U}, \bar{b})$ either do not intersect or coincide.*

PROOF: Suppose that the inclusion $\Phi(\mathfrak{U}, \bar{a}) \cap \Phi(\mathfrak{U}, \bar{b}) \subset \Phi(\mathfrak{U}, \bar{a})$ is strict. Consider the sequence of collections $\vec{f}_k \in A^\omega$, $k \in \omega$, defined as follows:

$$\vec{f}_k(i) = \begin{cases} \bar{a} \hat{\wedge} \bar{b} & \text{if } i \leq k \\ \bar{a} \hat{\wedge} \bar{a} & \text{if } i > k \end{cases}.$$

Consider the formula

$$\chi(\bar{y}, \bar{z}) = \forall x (\Psi(x, \bar{y}) \rightarrow \Psi(x, \bar{z}))$$

where $\Psi(x, \bar{y})$ is $\Phi(x, \bar{y}^1) \wedge \Phi(x, \bar{y}^2)$. It is obvious that

$$\mathfrak{U}^a \models \chi(\vec{f}_k, \vec{f}_m) \Leftrightarrow m \leq k.$$

Thus, $\text{Th}(M_\infty)$ has the order property, and, therefore, is unstable (see SHELAH, 1978). Lemma 3.3 is proved.

From Lemma 3.3 it follows that if $\text{Th}(M_\infty)$ is stable, then for any \exists^+ -formula $\Phi(x, \bar{z})$ the formula

$$\Psi(x, y) = \exists \bar{z} (\Phi(x, \bar{z}) \wedge \Phi(y, \bar{z}))$$

defines the equivalence in any M -system \mathfrak{A} on the set $\exists \bar{z} \Phi(\mathfrak{A}, \bar{z})$. In particular, for any $\bar{a} \in A$ the set $\Phi(\mathfrak{A}, \bar{a})$ is defined by any of its elements $b \in \Phi(\mathfrak{A}, \bar{a})$ and equals $\Psi(\mathfrak{A}, b)$.

\exists^+ -formula $\Phi(x, y)$ is said to be an \exists^+ -equivalence if for any M -system \mathfrak{A} the set $\{\langle a, b \rangle \mid \mathfrak{A} \models \Phi(a, b)\}$ is the equivalence on the set $\exists y \Phi(\mathfrak{A}, y)$. An \exists^+ -equivalence $\Phi(x, y)$ is called total if $M \models \forall x \Phi(x, x)$.

An \exists^+ -equivalence $\Phi(x, y)$ is called degenerate if

$$M \models \forall x \forall y (\Phi(x, y) \rightarrow x = y)$$

and unit if

$$M \models \forall x \forall y \Phi(x, y).$$

If $\Phi(x, y)$ is an \exists^+ -equivalence and $t(u, \bar{z})$ is a term, then \exists^+ -formula

$$\Psi(x, y) = \exists \bar{z} \exists u \exists v (\Phi(u, v) \wedge t(u, \bar{z}) = x \wedge t(v, \bar{z}) = y)$$

is called $t(u, \bar{z})$ -satellite of \exists^+ -equivalence $\Phi(x, y)$.

LEMMA 3.4. If $\text{Th}(M_\infty)$ is stable and $\Psi(x, y)$ is a satellite of \exists^+ -equivalence $\Phi(x, y)$, then $\Psi(x, y)$ is \exists^+ -equivalence.

PROOF: The immediate consequence of Lemma 3.3.

\exists^+ -formula

$$\exists u_0 \dots \exists u_n (x = u_0 \wedge y = u_n \wedge \bigwedge_{i=1}^n \Phi_i(u_{i-1}, u_i))$$

is called the combination of \exists^+ -formulae $\Phi_1(x, y), \dots, \Phi_n(x, y)$. Note that for $n = 1$ the combination of $\Phi_1(x, y)$ is equivalent to $\Phi_1(x, y)$.

LEMMA 3.5. Let $\text{Th}(M_\infty)$ be stable, X some set of \exists^+ -equivalences with free variables x, y , and let η be a minimal congruence on the system \mathfrak{A} which contains the set $\{\langle a, b \rangle \mid \Phi(x, y) \in X, \mathfrak{A} \models \Phi(a, b)\}$. Then, for $\langle a, b \rangle \in \eta$, it is necessary and sufficient that there exists a combination $\Psi(x, y)$ of some satellites of formulae from X and $\mathfrak{A} \models \Psi(a, b)$.

PROOF: Sufficiency is easily proved by induction on the number of combinations of members, due to the definition of congruence. To prove the necessity one has only to verify that the set

$$\{\langle a, b \rangle \in A^2 \mid \mathfrak{U} \models \Psi(a, b) \text{ for some combination } \Psi \\ \text{of satellites of formulae from } X\}$$

is the congruence.

LEMMA 3.6. *Let $\text{Th}(M_\infty)$ be stable. Then for any non-degenerate \exists^+ -equivalence $\Phi(x, y)$ there exist satellites $\Psi_1(x, y), \dots, \Psi_k(x, y)$ of the formula $\Phi(x, y)$ such that the combination of these satellites is a unit \exists^+ -equivalence.*

PROOF: By induction on n we construct the sets X_n , $1 < n < \omega$, of \exists^+ -formulae. As X_0 we take the set of all possible combinations of satellites of $\Phi(x, y)$. Let X_n have been already constructed. As Y_n we take the set of all possible simultaneous substitutions of formulae $\Psi(t_1, t_2)$ where $\Psi \in X_n$ to the formula Φ instead of atomic subformulae $t_1 = t_2$. Then as X_{n+1} we take the set of all possible combinations of satellites of formulae from $X_n \cup Y_n$. For any M -system \mathfrak{U} we define, by induction, the increasing sequence of congruences η_m , $m \in \omega$. As η_0 we take the zero congruence $\{\langle a, a \rangle \mid a \in A\}$. If η_m has already been constructed, and $h_m: \mathfrak{U} \rightarrow \mathfrak{U}/\eta_m$ is a natural homomorphism, then, as η_{m+1} we take the minimal congruence on \mathfrak{U} containing the set

$$\eta_m \cup \{\langle a, b \rangle \mid \mathfrak{U}/\eta_m \models \Phi(ha, hb)\}.$$

Making use of Lemma 3.5, one can easily show that for $1 < m < \omega$

$$\langle a, b \rangle \in \eta_{m+1} \Leftrightarrow (\mathfrak{U} \models \Psi(a, b) \text{ for some } \Psi(x, y) \in X_m).$$

Consider the congruence $\eta_\omega = \bigcup_{m \in \omega} \eta_m$. From the construction of η_m , $m \in \omega$, it easily follows that

$$\mathfrak{U} \models \forall x \forall y (\Phi(x, y) \rightarrow x = y).$$

From the completeness of M , closedness of M relative to Cartesian powers, and from non-degeneracy of $\Phi(x, y)$ we obtain that η_ω is the unit congruence of \mathfrak{U} . Let \mathfrak{U} be a M -system with a free pair $\langle e_1, e_2 \rangle$ in \mathfrak{U} (PALYUTIN, 1979). Then, according to Lemma 3.5 and the construction of η_m , $m \in \omega$, there exist $n \in \omega$ and $\Psi(x, y) \in X_n$ such that $\mathfrak{U} \models \Psi(e_1, e_2)$. From the freedom of $\langle e_1, e_2 \rangle$ in \mathfrak{U} and completeness of M we obtain $\mathfrak{B} \models \forall x \forall y \Psi(x, y)$

for any M -system \mathfrak{B} . Consequently, η_{n+1} is the unit congruence. Let n_0 be such minimal m for which \mathfrak{U}/η_m is the unit system. Then $n_0 > 0$ and let $\mathfrak{B} = \mathfrak{U}/\eta_{n_0-1}$. Since for any $\chi(x, y) \in X_{n_0-2}$ we have

$$\mathfrak{U} \models \chi(a, b) \Rightarrow \eta_{n_0-1}a = \eta_{n_0-1}b,$$

it follows that $\Psi(x, y)$ is equivalent in \mathfrak{B} to the combination of satellites $\Psi_1(x, y), \dots, \Psi_k(x, y)$ of the formula $\Phi(x, y)$. By the completeness of M Lemma 3.6 follows.

LEMMA 3.7. *Let $\text{Th}(M_\infty)$ be stable, let $\Phi(x, y)$ be non-degenerate ${}^+\exists$ -equivalence, and let for any M_+ -system \mathfrak{U} all not one-element sets $\Phi(\mathfrak{U}, \bar{a})$, $\bar{a} \in A$, be strongly \exists^+ -minimal. Then the variety M is categorical.*

PROOF: Due to Theorem 3.1, it is sufficient to show that in any M -system does not exist a strictly decreasing sequence of \exists^+ -sets.

Obviously, if $X \subseteq A$ is a one-element or strongly \exists^+ -minimal in \mathfrak{U} set, then for any term $t(x, \bar{y})$ and $\bar{a} \in A$ the set

$$\{b \mid \mathfrak{U} \models t(c, \bar{a}) = b \text{ for some } c \in X\}$$

is one-element or strongly minimal. Therefore, due to Lemma 3.6, it is sufficient to prove the following statement:

Let $\Psi(x, y)$ be total \exists^+ -equivalence and suppose that for any M -system \mathfrak{U} and any $a \in A$ the set $\Psi(\mathfrak{U}, a)$ does not include a strictly decreasing sequence of \exists^+ -sets in \mathfrak{U} . Let $\chi(x, \bar{y})$ be an \exists^+ -formula such that for any M -system \mathfrak{U} and any $\bar{a} \in A$ the set $\chi(\mathfrak{U}, \bar{a})$ is strongly \exists^+ -minimal or one-element. Then for any M -system \mathfrak{U} and any $\bar{a} \in A$ the set

$$Y = \{b \mid \mathfrak{U} \models \exists x(\chi(x, \bar{a}) \wedge \Psi(b, x))\}$$

does not include a strictly decreasing sequence of \exists^+ -sets in \mathfrak{U} .

Suppose this to be incorrect and Y to include a strictly decreasing sequence of \exists^+ -sets $\alpha_n(\mathfrak{U}, \bar{a}^n)$, $n \in \omega$. If for some $c_0 \in \chi(\mathfrak{U}, \bar{a})$ and $n_0 \in \omega$ the set $\Psi(\mathfrak{U}, c_0) \cap \alpha_{n_0}(\mathfrak{U}, \bar{a}^n)$ is empty, then the set

$$Z = \{c \mid \mathfrak{U} \models \chi(c, \bar{a}) \wedge \exists x(\Psi(x, c) \wedge \alpha_{n_0}(x, \bar{a}^{n_0}))\}$$

does not contain c_0 ; therefore, due to strong \exists^+ -minimality of $\chi(\mathfrak{U}, \bar{a})$, Z equals $\{c_1\}$ for some $c_1 \in \chi(\mathfrak{U}, \bar{a})$. Then the set $\Psi(\mathfrak{U}, c_1)$ includes a strictly decreasing sequence of \exists -sets $\alpha_k(\mathfrak{U}, \bar{a}^k)$, $n_0 \leq k < \omega$. This contradicts the supposition. So we have $\Psi(\mathfrak{U}, c) \cap \alpha_n(\mathfrak{U}, \bar{a}^n) \neq \emptyset$ for any $c \in \chi(\mathfrak{U}, \bar{a})$ and $n \in \omega$. Consider M -system $\mathfrak{B} = \mathfrak{U}^{\chi(\mathfrak{U}, \bar{a})}$ and its diagonal element

$f: \chi(\mathfrak{U}, \bar{a}) \rightarrow A$, i.e. $f(c) = c$, $c \in \chi(\mathfrak{U}, \bar{a})$. The lemma will be proved if we show that the sets $\Psi(\mathfrak{B}, f) \cap \alpha_n(\mathfrak{B}, \bar{a}^n)$ strictly decrease. Since the subsets $\alpha_n(\mathfrak{U}, \bar{a}^n)$, $n \in \omega$, of the set Y strictly decrease, for any $n \in \omega$ there exist $c_n \in \chi(\mathfrak{U}, \bar{a})$ and $b_n \in \Psi(\mathfrak{U}, c_n)$ such that

$$\mathfrak{U} \models \alpha_n(b_n, \bar{a}^n) \wedge \neg \alpha_{n+1}(b_n, \bar{a}^{n+1}).$$

Choose $g_n \in B$ with the following properties:

$$g_n(c_n) = b_n, \quad g_n(c) \in \alpha_n(\mathfrak{U}, \bar{a}^n) \cap \Psi(\mathfrak{U}, c) \quad \text{for } c \in \chi(\mathfrak{U}, \bar{a}) \setminus \{c_n\}.$$

It is clear that

$$\mathfrak{U} \models \Psi(g_n, f) \wedge \alpha_n(g_n, \bar{a}^n) \wedge \neg \alpha_{n+1}(g_n, \bar{a}^{n+1}).$$

Lemma 3.7 is proved.

PROOF OF THEOREM 3.2. The implication $(1) \Rightarrow (2)$ was proved by M. Morley (see CHANG and KEISLER, 1973). Suppose that M is not categorical and $\text{Th}(M_\omega)$ is superstable. By induction on $n \in \omega$ we construct a total non-degenerate \exists^+ -equivalences $\Phi_n(x, y)$, $n \in \omega$, with the following properties:

- (a) $M \models \forall x \forall y (\Phi_{n+1}(x, y) \rightarrow \Phi_n(x, y))$, $n \in \omega$;
- (b) $M \models \exists x \exists y (\Phi_n(x, y) \wedge \neg \Phi_{n+1}(x, y))$.

As $\Phi_0(x, y)$ we take the formula $x = x \wedge y = y$. Let $\Phi_n(x, y)$ have already been constructed. Due to Lemma 3.7 there exist an M -system \mathfrak{U} and $a_0 \in A$ for which the set $\Phi_n(\mathfrak{U}, \bar{a}_0)$ is not one-element and not strongly \exists^+ -minimal. Therefore, there exist such an \exists^+ -formula $\Psi(x, \bar{z})$ and $\bar{b} \in A$ that the set $\Phi_n(\mathfrak{U}, a_0) \cap \Psi(\mathfrak{U}, \bar{b})$ contains more than one element and does not coincide with $\Phi_n(\mathfrak{U}, a_0)$. We shall suppose that $a_0 \in \Psi(\mathfrak{U}, \bar{b})$.

Consider the formula

$$\Phi(x, y) = (\Phi_n(x, y) \wedge \exists \bar{z} (\Psi(x, \bar{z}) \wedge \Psi(y, \bar{z}))).$$

According to Lemma 3.3, $\Phi(x, y)$ is \exists^+ -equivalence. Since

$$\Phi_n(\mathfrak{U}, a_0) \cap \Psi(\mathfrak{U}, \bar{b})$$

contains more than one element, it follows that $\Phi(x, y)$ is non-degenerate. If $\Phi(x, y)$ is total, then as Φ_{n+1} we take Φ . If $\Phi(x, y)$ is not total, then we consider $t_i(x, \bar{u})$ -satellite $\Psi_i(x, y)$ of the formula Φ from Lemma 3.6 where $i \in \{1, \dots, k\}$ is the minimal number for which $\Psi_i(x, y)$ is a non-degenerate \exists^+ -equivalence. Since the combination of $\Psi_i(x, y), \Psi_{i+1}(x, y), \dots, \Psi_k(x, y)$

is a unit equivalence, $\Psi_i(x, y)$ is the total \exists^+ -equivalence. Take $a_1 \in A$ and $\bar{c} \in A$ for which the set

$$t_i(\Phi(\mathfrak{U}, a_1), \bar{c}) = \{a \mid \mathfrak{U} \models \exists x(\Phi(x, a_1) \wedge t_i(x, \bar{c})) = a\}$$

contains more than one element. Consider M -system $\mathfrak{B} = \mathfrak{U} \times \mathfrak{U}$. Let $b_0 = \langle a_0, a_1 \rangle$. Then the set $t_i(\Phi(\mathfrak{B}, b_0), c)$ contains more than one element, and $\Phi(\mathfrak{B}, b_0)$ is strictly contained in $\Phi_n(\mathfrak{B}, b_0)$. If there exist two unequal elements $a_1, a_2 \in \Phi_n(\mathfrak{B}, b_0)$ for which $\mathfrak{U} \models t_i(a_1, \bar{c}) = t_i(a_2, \bar{c})$, then by Lemma 3.3, as Φ_{n+1} we can take the formula

$$\Phi_n(x, y) \wedge \exists \bar{u}(t_i(x, \bar{u}) = t_i(y, \bar{u})).$$

If such a_1, a_2 do not exist, then, due to the totality of $\Psi_i(x, y)$ and Lemma 3.3, as Φ_{n+1} we can take the formula

$$\Phi_n(x, y) \wedge \exists \bar{u}(\Psi_i(t_i(x, \bar{u}), t_i(y, \bar{u}))).$$

Let the signature Σ' be obtained by adding to Σ the constants c_δ for any finite sequence $\delta \in \omega^{<\omega}$ of natural numbers.

It is known (SHELAH, 1978) that we arrive at a contradiction with superstability if we show the consistency with $\text{Th}(M_\infty)$ of the following set of propositions:

- (1) $\Phi_n(c_{\langle k_1, \dots, k_n \rangle}, c_{\langle k_1, \dots, k_n, m \rangle})$, $n, m, k_1, \dots, k_n \in \omega$;
- (2) $\neg \Phi_{n+1}(c_{\langle k_1, \dots, k_n, m \rangle}, c_{\langle k_1, \dots, k_n, 1 \rangle})$, $n, m, l, k_1, \dots, k_n \in \omega$, $m \neq l$.

Let \mathfrak{U} be an M_∞ -system. From conditions (a), (b) on the total \exists^+ -equivalences Φ_n , $n \in \omega$, it follows that for any $n \in \omega$ there exist elements $a_n, a'_n \in A$ for which

- (a) $\mathfrak{U} \models \Phi_n(a_n, a'_n)$,
- (b) $\mathfrak{U} \models \neg \Phi_{n+1}(a_n, a'_n)$.

If δ is a finite sequence of natural numbers, then by $l(\delta)$ we denote the length of δ . For any finite sequence of natural numbers δ, γ , $l(\gamma) \neq 0$ we define the element c_δ^γ as follows:

$$c_\delta^\gamma = \begin{cases} a'_{l(\gamma)-1} & \text{if } \gamma \text{ is the initial segment of } \delta, \\ a_{l(\gamma)-1} & \text{otherwise.} \end{cases}$$

Let S be a set of all non-empty finite sequences. Consider an M -system \mathfrak{U}^S . It is easy to verify that formulae (1), (2) will be true in enriching \mathfrak{U}^S up to

the signature Σ' if we take, as c_δ , the following functions: $c_\delta(\gamma) = c_\delta^\gamma$, $\gamma \in S$. Theorem 3.2 is proved.

Remark. Note that for our proof of Theorem 3.2 only the following conditions on the class M are necessary:

- (a) M is closed relative to homomorphic images and Cartesian powers;
- (b) M is complete with respect to strictly multiplicative stable $\forall\exists$ -formulae.

4. Countable spectrum

In this section we complete the proof of Theorem 1, having proved the following statement:

THEOREM 4.1. *If M is a complete variety, the following conditions are equivalent:*

- (1) M is non-categorical,
- (2) there exist 2^ω isomorphism types of countable M -systems,
- (3) for some $n \in \omega$ there exist $2^\omega n$ -types of signature Σ which are consistent with $\text{Th}(M_\infty)$.

PROOF: The implication (3) \Rightarrow (2) is obvious. The implication (2) \Rightarrow (1) has been proved by M. Morley (cf. CHANG and KEISLER, 1973). We prove (1) \Rightarrow (3). Let a variety M be complete and not categorical. Consider a free M -system \mathfrak{U} with free generators e_1, e_2 . Due to Theorem 2.1, \mathfrak{U} is infinite.

Note some properties of M_∞ -system \mathfrak{U} .

- (a) If \exists^+ -set X in \mathfrak{U} contains more than one element, then X is infinite.
- (b) Let $\Phi(x, \bar{y})$ be \exists^+ -formula, X an \exists^+ -set in \mathfrak{U} , $\bar{b}^1, \bar{b}^2 \in A$, \exists^+ -sets $\Phi(\mathfrak{U}, \bar{b}^1)$ and $\Phi(\mathfrak{U}, \bar{b}^2)$ be different non-empty subsets of X , and $\Phi(\mathfrak{U}, \bar{b}^1)$ contain more than one element. Then there exists $\bar{b}^3 \in A$ such that $\Phi(\mathfrak{U}, \bar{b}^3)$ is a distinct from $\Phi(\mathfrak{U}, \bar{b}^1)$ subset of X and contains more than one element.
- (c) Let $\Psi(x, \bar{y})$ be \exists^+ -formula, $\bar{a}, \bar{b} \in A$ and $\Psi(\mathfrak{U}, \bar{a})$ a proper non-empty subset $\Psi(\mathfrak{U}, \bar{b})$. Then there exists $\bar{c} \in A$ such that $\Psi(\mathfrak{U}, \bar{a})$ is strictly contained in $\Psi(\mathfrak{U}, \bar{c})$ and $\Psi(\mathfrak{U}, \bar{c})$ is strictly contained in $\Psi(\mathfrak{U}, \bar{b})$.

Due to the completeness of M , to prove these properties it is sufficient to obtain some M_∞ -systems $\mathfrak{B}_1, \mathfrak{B}_2, \mathfrak{B}_3$ in which (a), (b) and (c) are satisfied, respectively. It is clear that as \mathfrak{B}_1 we can take M_∞ -system $\bigcup_{n \in \omega} \mathfrak{U}_n$

where $\mathfrak{U}_0 = \mathfrak{U}$ and $\mathfrak{U}_{n+1} = \mathfrak{U}_n \times \mathfrak{U}_n$, $n \in \omega$. As \mathfrak{B}_2 and \mathfrak{B}_3 we can take

the reduced power of \mathfrak{A} modulo F , where F is the Fréchet filter on ω ($F = \{N \subseteq \omega \mid \omega \setminus N \text{ is finite}\}$). We verify on $\mathfrak{B} = \mathfrak{A}^\omega / F$ property (b). Verification of (c) is done in a similar way. Let $X = \Psi(\mathfrak{B}, \bar{f}/F)$, $\Phi(\mathfrak{B}, \bar{g}^1/F) \subseteq X$, $\Phi(\mathfrak{B}, \bar{g}^2/F) \subseteq X$, $\Phi(\mathfrak{B}, \bar{g}^2/F)$ be non-empty and let $\Phi(\mathfrak{B}, \bar{g}^1/F)$ be distinct from $\Phi(\mathfrak{B}, \bar{g}^2/F)$ and contain more than one element. We may suppose that $|\Phi(\mathfrak{B}, \bar{g}^2/F)| = 1$. Due to the known properties of the reduced power, there exists an infinite set $N_0 \subseteq \omega$ such that for any $i \in N_0$ we have that $\Phi(\mathfrak{A}, \bar{g}^1(i))$, $\Phi(\mathfrak{A}, \bar{g}^2(i))$ are non-empty different subsets of $\Psi(\mathfrak{A}, \bar{f}(i))$, and $\Phi(\mathfrak{A}, \bar{g}^1(i))$ contains more than one element. Take an infinite subset $N_1 \subseteq N_0$ for which the set $N_0 \setminus N_1$ is also infinite. Then, as the required $\bar{b}^3 \in B$ we can take \bar{g}^3/F , where \bar{g}^3 is defined as follows:

$$\bar{g}^3(i) = \begin{cases} \bar{g}^1(i) & \text{if } i \in N_1, \\ \bar{g}^2(i) & \text{if } i \in \omega \setminus N_1. \end{cases}$$

Case 1. The conclusion of Lemma 3.3 does not hold.

Then there exist \exists^+ -formula $\Psi(\bar{x}, \bar{y})$ and $\bar{a}, \bar{b} \in A$ such that $\Psi(\mathfrak{A}, \bar{a})$ is non-empty proper subset of $\Psi(\mathfrak{A}, \bar{b})$. From property (c) it follows that there exists a set of term collections $\{\bar{t}_r(x, y) \mid r \in Q\}$ where Q is the set of rational numbers for which the following holds:

- (d) If $r_1 < r_2$, then \exists^+ -set $\Phi(\mathfrak{A}, \bar{t}_{r_1}(e_1, e_2))$ is strictly contained in \exists^+ -set $\Phi(\mathfrak{A}, \bar{t}_{r_2}(e_1, e_2))$.

Let $q(u_1, u_2)$ be 2-type which is realized in \mathfrak{A} by the pair $\langle e_1, e_2 \rangle$. It is clear that for any initial segment $Z \subseteq Q$ a set of formulae

$$\begin{aligned} X(Z) = q(u_1, u_2) \cup \{ \forall y (\Phi(y, \bar{t}_r(u_1, u_2)) \rightarrow \Phi(y, \bar{x})) \mid r \in Z \} \cup \\ \cup \{ \forall y (\Phi(y, x) \rightarrow \Phi(y, \bar{t}_r(u_1, u_2))) \mid r \in Q \setminus Z \} \end{aligned}$$

with free variables $u_1, u_2, x_1, \dots, x_n$ is consistent with $\text{Th}(M_\infty)$. It is clear that for the initial segments $Z_1 \neq Z_2$ the set $X(Z_1) \cup X(Z_2) \cup \text{Th}(M_\infty)$ is inconsistent. Thus, there exist 2^ω $(n+2)$ -types consistent with $\text{Th}(M_\infty)$.

Case 2. The conclusion of Lemma 3.3 is true.

Note that in Lemmas 3.4–3.7 we used the stability of $\text{Th}(M_\infty)$ only to be able to apply Lemma 3.3. For that reason in this case those lemmas and their corollaries can be applied.

Having repeated the arguments from the proof of Theorem 3.2 at the construction of \exists^+ -formula Φ_{n+1} , we can prove the following property:

(d) For any non-degenerate \exists^+ -equivalence $\Psi(x, y)$ there exists a non-degenerate \exists^+ -equivalence $\chi(x, y)$ for which

$$M \models \forall x(\chi(x, x) \leftrightarrow \Psi(x, x)), \quad M \models \forall x \forall y(\chi(x, y) \rightarrow \Psi(x, y))$$

and

$$M \models \exists x \exists y(\Psi(x, y) \wedge \neg \chi(x, y)).$$

Suppose that in some M_∞ -system there exist two different elements, each of them being definable by some \exists^+ -formula. Then there exists an M_∞ -system \mathfrak{U}_0 whose all elements are definable with the help of \exists^+ -formula. Then, due to (d) and (b) we obtain that any infinite set definable by \exists^+ -formula contains two non-intersecting infinite subsets definable with the help of \exists^+ -formulae. Consequently, there exist 2^ω 1-types of the signature Σ which are consistent with $\text{Th}(M_\infty)$. Thus, we can assume that if $\Phi(x)$ and $\Psi(x)$ are \exists^+ -formulae, and for M_∞ -system \mathfrak{U} $\Phi(\mathfrak{U})$ and $\Psi(\mathfrak{U})$ are one-element, then $\Phi(\mathfrak{U}) = \Psi(\mathfrak{U})$. Below we shall make use of this assumption.

Subcase 2a. There exists \exists^+ -formula $\Phi_0(x)$ such that for any M_∞ -system \mathfrak{U} the set $\Phi_0(\mathfrak{U})$ is infinite, and for any \exists^+ -formula $\Psi(x)$ the set $\Phi_0(\mathfrak{U}) \cap \Psi(\mathfrak{U})$ is empty, one-element or coincides with $\Phi_0(\mathfrak{U})$.

We prove that under these conditions the following fact takes place.

(e) If $\Phi(x, y)$, $\Psi(x, y)$ are \exists^+ -equivalences,

$$M \models \forall x(\Phi(x, x) \leftrightarrow \Phi_0(x)), \quad M \models \forall x \forall y(\Psi(x, y) \rightarrow \Phi(x, y)),$$

$$M \models \exists x \exists y(\Phi(x, y) \wedge \neg \Psi(x, y)), \quad \Phi(\mathfrak{U}, a_0) = \Psi(\mathfrak{U}, a_0)$$

for some M_∞ -system \mathfrak{U} and $a_0 \in \Phi_0(\mathfrak{U})$, then a_0 is defined in \mathfrak{U} by some \exists^+ -formula.

Consider the set

$$\text{Th}(M_\infty) \cup \mathcal{D}^+(\mathfrak{U}) \cup \{\exists y(\Phi(y, a_0) \wedge \neg \Psi(y, a_0))\} \quad (*)$$

where $\mathcal{D}^+(A)$ is the positive diagram of \mathfrak{U} . Suppose that this set is non-consistent. Then, for some \exists^+ -formula $\chi(x)$, we have $\mathfrak{U} \models \chi(a_0)$ and

$$\text{Th}(M_\infty), \chi(x) \vdash \forall y(\Phi(y, x) \rightarrow \Psi(y, x)).$$

From the properties of formulae Φ_0 , Φ , and Ψ we obtain that a_0 is defined in \mathfrak{U} with the help of \exists^+ -formula $\chi(x) \wedge \Phi_0(x)$. Let now a_0 be defined in \mathfrak{U} by no \exists^+ -formula. Then the set (*) is consistent, whence there exist an M_∞ -system \mathfrak{B} and homomorphism $h: \mathfrak{U} \rightarrow \mathfrak{B}$ for which $\Phi(\mathfrak{B}, ha_0)$

$\neq \Psi(\mathfrak{B}, ha_0)$. Consider the mapping $f: A \rightarrow A \times B$ which associates to an element a the pair $\langle a, ha_0 \rangle$. It is clear that f isomorphically embeds \mathfrak{A} into $\mathfrak{A} \times \mathfrak{B}$, $\Phi(\mathfrak{A} \times \mathfrak{B}, fa_0) \neq \Psi(\mathfrak{A} \times \mathfrak{B}, fa_0)$ and for any $a, b \in A$ from $\mathfrak{A} \models \neg \Psi(a, b)$ there follows $\mathfrak{A} \times \mathfrak{B} \models \neg \Psi(fa, fb)$. Having constructed the tower of such extensions, we can obtain M_∞ -system \mathfrak{A}^* for which $\Phi(\mathfrak{A}^*, a) \neq \Psi(\mathfrak{A}^*, a)$ for any element $a \in \Phi_0(\mathfrak{A}^*)$ which is non-definable in \mathfrak{A}^* with the help of \exists^+ -formula. Thus, by the completeness of M , property (e) follows.

If follows from properties (d) and (e) that there exists a sequence of \exists^+ -equivalences $\Psi_n(x, y)$, $n \in \omega$, with the following properties:

- (1) $M \models \forall x (\Psi_n(x, x) \leftrightarrow \Phi_0(x))$, $n \in \omega$;
- (2) $M \models \forall x \forall y (\Psi_{n+1}(x, y) \rightarrow \Psi_n(x, y))$, $n \in \omega$;
- (3) in every M_∞ -system \mathfrak{A} , any Ψ_n -class which does not contain an element definable with the help of \exists^+ -formula contains more than one Ψ_{n+1} -class.

Since there exist 2-generated M_∞ -systems, it follows from properties (1)–(3) that there are 2^ω 3-types of the signature Σ which are compatible with $\text{Th}(M_\infty)$.

Subcase 2b. Negation of Subcase 2a.

A set definable in the system \mathfrak{A} with the help of some \exists^+ -formula $\Phi(x)$ is called *absolute \exists^+ -set in \mathfrak{A}* . If every infinite \exists^+ -set in M_∞ -systems contains two non-intersecting infinite absolute \exists^+ -sets, then there exist 2^ω 1-types of the signature Σ which are compatible with $\text{Th}(M_\infty)$. Therefore, in the sequel we shall assume that there exists \exists^+ -formula $\Psi_0(x)$ such that for any M_∞ -system \mathfrak{A} any two infinite absolute \exists^+ -sets contained in $\Psi_0(\mathfrak{A})$ have infinite intersection.

Let $\{\Psi_n(x) | n \in \omega\}$ be such a sequence of \exists^+ -formulae that in any M_∞ -system \mathfrak{A} defines a strictly decreasing sequence of infinite absolute \exists^+ -sets, and for any \exists^+ -formula $\Psi(x)$ either $\Psi(\mathfrak{A}) \cap \Psi_0(\mathfrak{A})$ contains not more than one element, or for some $n \in \omega$ there holds the inclusion $\Psi_n(\mathfrak{A}) \subseteq \Psi(\mathfrak{A})$. Let the signature Σ_1 be obtained from Σ by adding up a new constant c and consider the theory T_1 which besides the axioms $\text{Th}(M_\infty)$ contains the axiom $\{\Psi_n(c) | n \in \omega\}$ and axiom $\Psi(c)$ as well if there exists \exists^+ -formula $\Psi(x)$ for which $\Psi(\mathfrak{A})$ is one-element, and $\Psi(\mathfrak{A}) \subseteq \Psi_n(\mathfrak{A})$ for any M_∞ -system \mathfrak{A} and any $n \in \omega$.

From the compactness theorem it follows that for any model \mathfrak{A} of the theory T_1 the set $\Psi_0(\mathfrak{A})$ does not contain disjunct absolute \exists^+ -sets containing more than one element.

Subsubcase 2b₁. There exists model \mathfrak{U} of the theory T_1 in which a subsystem \mathfrak{U}_0 , consisting of elements definable by \exists^+ -formulae of the signature Σ_1 , contains more than one element.

Let n_0 be a number such that an element definable in $\mathfrak{U}_0 \upharpoonright \Sigma$ by \exists^+ -formula of the signature Σ (if any) does not belong to $\Psi_{n_0}(\mathfrak{U}) \setminus \Psi_{n_0-1}(\mathfrak{U})$. If there is no such element, let $n_0 = 0$. We state that there exists an element $a_1 \in A_0$ for which $\mathfrak{U} \models \Psi_{n_0}(a_1) \wedge \neg \Psi_{n_0+1}(a_1)$. Indeed, if not so, M_ω -system $\mathfrak{U}_0 \upharpoonright \Sigma$ would have an extension $\mathfrak{U} \upharpoonright \Sigma$ which is an M -system and in which all the elements $\Psi_{n_0}(\mathfrak{U}_0)$ satisfy $\Psi_{n_0+1}(x)$. From the completeness of M , due to the theorem of compactness, we infer that there exists a tower

$$\mathfrak{B}_0 \subseteq \mathfrak{B}_1 \subseteq \dots \subseteq \mathfrak{B}_n \subseteq \dots, \quad n \in \omega,$$

of M -systems such that for any $n \in \omega$ we have $\Psi_{n_0}(\mathfrak{B}_n) \subseteq \Psi_{n_0+1}(\mathfrak{B}_{n+1})$. Then $\Psi_{n_0}(\mathfrak{B}_\omega) = \Psi_{n_0+1}(\mathfrak{B}_\omega)$, where $\mathfrak{B}_\omega = \bigcup_{n \in \omega} \mathfrak{B}_n$. This contradicts the completeness of M .

Hence and from the definition of the theory T_1 we obtain that there exists such a number m_0 for which the set $\Psi_{m_0}(\mathfrak{U})$ does not contain an element definable in \mathfrak{U} by \exists^+ -formula of the signature Σ . Let $\{a_1\} = \Phi(\mathfrak{U}, c)$, where $\Phi(x, y)$ is \exists^+ -formula of the signature Σ . The set

$$X = (\Psi_0(\mathfrak{U}) \cap \exists x(\Phi(\mathfrak{U}, x) \wedge \Psi_{m_0}(x)))$$

contains c since it contains the element a_1 which is defined in \mathfrak{U} by no \exists^+ -formula of the signature Σ . Then there exists $a_2 \in A$ for which $\mathfrak{U} \models \Phi(c, a_2) \wedge \Psi_{m_0}(a_2)$. The set

$$Y = (\Psi_0(\mathfrak{U}) \cap \exists x(\Phi(x, \mathfrak{U}) \wedge \Psi_{n_0+1}(x)))$$

does not contain c and contains a_2 , whence $Y = \{a_2\}$. This contradicts the condition $a_2 \in \Psi_{m_0}(\mathfrak{U})$ and the choice of m_0 .

Subsubcase 2b₂. Negation of Case 2b₁.

Let $\{\Psi'_n(x, y) \mid n \in \omega\}$ be a sequence of \exists^+ -formulae of the signature Σ and let \mathfrak{U} be a model of T_1 with the following properties:

- (a) $\Psi'_0(\mathfrak{U}, c) = \Psi_0(\mathfrak{U})$;
- (b) $\Psi'_n(c, c)$, $n \in \omega$;
- (c) $\Psi'_n(\mathfrak{U}, c)$ is infinite, $n \in \omega$;
- (d) $\Psi'_{n+1}(\mathfrak{U}, c) \subseteq \Psi'_n(\mathfrak{U}, c)$, $n \in \omega$;
- (e) for any \exists^+ -formula $\chi(x, y)$ of signature Σ either $\chi(\mathfrak{U}, c) \cap \Psi'_0(\mathfrak{U}, c)$ contains no more than one element or for some $n \in \omega$ there takes place the inclusion $\Psi'_n(\mathfrak{U}, c) \subseteq \chi(\mathfrak{U}, c)$.

Let a_0 be an element distinct from c and satisfying in \mathfrak{U} all $\Psi'_n(x, c)$, $n \in \omega$. Having repeated the construction of Subsubcase 2b₁ for the subsystem \mathfrak{U}_0 generated in \mathfrak{U} by the element a_0 , one can show that there exists an element $a_1 \in A_0$ for which

$$\mathfrak{U} \models \Psi_0(a_1) \wedge \neg \Psi_1(a_1).$$

Let $a_1 = t(a_0, c)$, where $t(x, y)$ is a term of the signature Σ .

Consider the set

$$X = \{a \in \Psi_0(\mathfrak{U}) \mid t(b, c) = a \text{ for some } b \in \Psi_0(\mathfrak{U})\}.$$

Since $X \cap \neg \Psi_1(\mathfrak{U}) \neq \emptyset$ we have $a_0 \in X$. Therefore, the set

$$Y = \{a \in \Psi_0(A) \mid \mathfrak{U} \models \Psi_1(t(a, c))\}$$

is non-empty. It does not contain a_0 , thus we have $Y = \{c\}$. Hence, due to $a_0 \in X$, we have $t(c, c) = a_0$. Thus, there exists more than one element definable in \mathfrak{U} by \exists^+ -formula. This contradicts the aforesaid assumption.

Theorem 4.1 is proved.

References

- ABAKUMOV, A. I., E. A. PALYUTIN, Yu. E. SHISHMAREV, M. A. TAITZLIN, 1972, *Categorical quasivarieties*, Algebra i Logika, vol. 11, No. 1
- CHANG, C. C., and H. J. KEISLER, 1973, *Model theory* (North-Holland Publishing Company)
- ERSHOV, Yu. L., and E. A. PALYUTIN, 1979, *Mathematical logic* (Nauka, Moscow)
- LACHLAN, A., 1972, *Complete varieties of algebras*, NAMS, vol. 19, No 5
- PALYUTIN, E. A., 1973, *On the spectrum of complete quasivarieties*, Proceedings of the XII All-Union Algebraic Conference, Sverdlovsk
- PALYUTIN, E. A., 1975, *Description of categorical quasivarieties*, Algebra i Logika, vol. 14, No. 2
- PALYUTIN, E. A., 1979, *On categorical positive Horn classes*, Algebra i Logika, vol. 18, No. 1
- SHELAH, S., 1978, *Classification theory* (North-Holland Publishing Company)

ON ALGEBRAICALLY CLOSED MODELS OF THEORIES OF COMMUTATIVE RINGS

JOACHIM REINEKE

Institute of Mathematics, Hannover University, Hannover, F.R.G.

Let T_0 be the theory of all commutative rings with identity. Let T be an arbitrary theory in a countable language L , containing the language of rings with identity. T is called a *special theory* if

- (i) $T_0 \subset T$,
- (ii) if R is a model of T , $a \in R$, then $R[x]/(ax)$ is a model of T ,
- (iii) $T = T_\forall$.

In § 1 we will present examples of special theories. Then in § 2 we will prove some theorems about algebraically closed models of special theories. Finally, in § 3, we have some of our main results as the following:

THEOREM. *Let T be a special theory. Then T has no model companion.*

THEOREM. *Let T be a special theory. Then there is an A_3 -sentence which holds in all finitely generic models of T and whose negation holds in all infinitely generic models of T .*

0. Preliminaries

For the model-theoretic background one should consult BARWISE and ROBINSON (1970). For ring-theoretical details and notions used in this paper see NAGATA (1962). By ‘ring’ we will always mean a commutative ring with identity. If R is a ring and if A is an ideal of R , then $\text{rad}(A)$ denotes the prime-radical of A and $J(A)$ the Jacobson radical of A . $R[\bar{x}]$ denotes the polynomial ring of R in a finite set of variables \bar{x} .

If $A \subset R$, then by (A) we denote the ideal generated by A and $\text{Ann}(A)$ is the ideal of elements of R which annihilates every element of A .

1. Examples of special theories

Clearly, T_0 , the theory of all commutative rings with identity, is a special theory. So we will have as a corollary of our main theorems the theorems of CHERLIN (1973). We will give further examples of special theories.

DEFINITION. Let R be a ring. Then we denote by $B(R)$ the Boolean algebra of idempotents of R .

LEMMA 1. Let R be a ring and let a be an element of R . Then $B(R)$ is canonically isomorphic to $B(R[x]/(ax))$.

PROOF: Define $h: B(R) \rightarrow B(R[x]/(ax))$ as follows: $h(e) := e + (ax)$. Clearly, h is an injective homomorphism of Boolean algebras. We have to show that h is surjective. Let $f(x) = \sum_i b_i x^i$ be an element of $R[x]$ and assume that $f^2 - f = p \cdot ax$ for some polynomial $p \in R[x]$. By comparison of coefficients we get $b_0^2 - b_0 = 0$. We will show that $h(b_0) = f + (ax)$.

Again by comparison of coefficients we have that $b_1(2b_0 - 1)$ is an element of (a) . Since b_0 is an idempotent, $2b_0 - 1$ is a unit of R . Hence $b_1 \in (a)$. By induction we can assume that b_1, \dots, b_k are elements of (a) . Hence we get $b_{k+1}(2b_0 - 1) \in (a)$. Therefore, $b_{k+1} \in (a)$. We can conclude that $b_1, \dots, b_n \in (a)$ and $h(b_0) = b_0 + (ax) = f + (ax)$. This proves our lemma.

Now, let I be the canonical interpretation with the following property: For every ring and every formula x of the language of Boolean algebras:

$$B(R) \models X \text{ if and only if } R \models X^I.$$

If F is a set of formulas of the language of Boolean algebras, then $F^I := \{X^I; X \in F\}$. Hence $B(R) \models K$ if and only if $R \models K^I$.

COROLLARY 2. Let K be a universal theory of Boolean algebras. Let $T(K) := T_0 \cup K^I$. Then $T(K)$ is a special theory.

PROOF: This follows immediately from Lemma 1. Clearly, if K is the theory of all Boolean algebras, then $T(K) = T_0$ is the theory of all commutative rings.

EXAMPLE 1. (a) Let $K = \{\forall a(a = 0 \vee a = 1)\}$. Then $T(K)$ is the theory of all indecomposable rings, i.e. rings without non-trivial idempotents. From Corollary 2 we can conclude that the theory of all indecomposable rings is a special theory. So we will have as a corollary of our main theorems the theorems of PODEWSKI and REINEKE (to appear) as well.

(b) Let B be a finite Boolean algebra and let n be the number of atoms

of B . Let L be the language of L_0 enriched by new constants \mathbf{b} for each $b \in B$. Let

$$K = \{\forall a \bigvee_{b \in B} a = b\}.$$

From Corollary 2 we can conclude that $T(K)$ is a special theory. Clearly, R is a model of $T(K)$ if and only if R is isomorphic to a finite product of at most n indecomposable rings.

We will now give another class of examples of special theories.

DEFINITION. Let L be a countable language containing the language of ring theory. Further, let H be the set of all terms with no free variables. Then define

$$H[x] := \{t_0 + t_1 x + \dots + t_n x^n; t_i \in H, n \geq 1\}$$

with x as a new free variable.

DEFINITION. Let T be a theory in a language L , let $f(x)$ be an element of $H[x]$, and let $b_0, \dots, b_n \in H$. Then we will call T *axiomatizable by f and b_0, \dots, b_n* if

- (i) $T_0 \subset T$,
- (ii) for all rings R , R is a model of T if and only if

$$R \models \forall x(f(x) = 0 \rightarrow \bigvee_{i=0}^n x = b_i).$$

LEMMA 2. Let R be a ring and $f(x) \in R[x]$ and let b_0, \dots, b_k be elements of R . Suppose that in R holds:

$$R \models \forall x(f(x) = 0 \rightarrow \bigvee_{i=0}^k x = b_i);$$

and further we assume that, for all $i = 0, \dots, k$, $f'(b_i)$ is a unit in R (where f' is the derivation of f). Then for all $a \in R$

$$R[x]/(ax) \models \forall x(f(x) = 0 \rightarrow \bigvee_{i=0}^k x = b_i).$$

PROOF: Let $f(x) = \sum_{i=0}^n a_i x^i$ and let $g(x) = \sum_{i=0}^r g_i x^i$ be a polynomial

of $R[x]$ such that $f(g(x) + (ax)) = (ax)$. Then there exists a polynomial $p \in R[x]$ such that $f(g(x)) = pax$. Therefore, $f(g_0) = 0$. It follows that $f'(g_0)$ is a unit in R .

By comparison of coefficients we have

$$\sum_{i=1}^n ia_i g^{i-1} g_1 = f'(g_0) g_1 \in (a).$$

Therefore $g_1 \in (a)$. By induction we can assume that $g_1, \dots, g_i \in (a)$. Then $g = g_0 + x^{i+1}(g_{i+1} + xh) + qax$ for some $h, q \in R[x]$. Again by comparison of coefficients in $f(g(x)) = pax$ we get $\sum_{j=1}^n ja_j g_0^{j-1} g_{i+1} \in (a)$. Therefore $g_{i+1} f'(g_0)$ is an element of (a) . It follows that g_{i+1} is an element of (a) . From this we can conclude that $g(x) - g_0 \in (ax)$, i.e. $R[x]/(ax) \models g(x) = g_0$. This proves our lemma.

COROLLARY 3. Let $f(x) \in H[x]$ and $b_0, \dots, b_n \in H$, $c_0, \dots, c_n \in H$ and let T be axiomatizable by f and b_0, \dots, b_n . Further assume that

$$T \vdash \bigwedge_{i=0}^n f'(b_i) \cdot c_i = 1.$$

Then T is a special theory.

PROOF: The proof follows from Lemma 2.

EXAMPLE 2. (a) Let $f(x) = x$. Then $f'(0) = 1$ and T_f is the theory of all commutative rings.

(b) The theory of indecomposable rings: let $f(x) = x^2 - x$. Then $f'(x) = 2x - 1$ and $f'(0) = -1$ and $f'(1) = 1$ are units and again we can conclude from Corollary 3 that the theory T_f is a special theory.

(c) Let Q be the field of rational numbers. For all $n \geq 1$ we define: $T_n := D(Q) \cup T_0 \cup \{\forall x(x^{2^n} - x = 0 \rightarrow x = 0 \vee x = 1)\}$ where $D(Q)$ is the diagram of Q . Let $f(x) = x^{2^n} - x$. Then $f'(0) = -1$ and $f'(1) = 2n - 1$ are units for all models of T_n . Clearly, Q is a model of T_n ; but every algebraically closed field is not a model of T_n ($n \geq 2$).

From Corollary 3 we can conclude that for all $n \geq 1$ T_n is a special theory.

(d) Let $a_1, \dots, a_n \in Q$, $a_i \neq a_j$ and $f(x) = \prod_{i=1}^n (x - a_i)$. Then $f'(a_i) \neq 0$ for all i .

Let

$$T_f := D(Q) \cup T_0 \{\forall x(f(x) = 0 \rightarrow \bigvee_{i=1}^n x = a_i)\}.$$

From Corollary 3 it follows that T_f is a special theory.

Similarly, one can give more examples of special theories which are axiomatizable by the roots of a polynomial. We will leave this to the reader.

We will show in the next section that special theories have no model companion. Now Lipschitz and Saracino and Carson have shown that the theory T_0^{red} of all commutative rings without nilpotent elements has a model companion (LIPSCHITZ and SARACINO, 1973).

Clearly, T_0^{red} is axiomatizable by $f(x) = x^2$. But $f'(0) = 2 \cdot 0 = 0$. Therefore we really have to assume that $f'(b)$ are units in order to prove that theories axiomatizable by a polynomial and the roots of this polynomial do not have a model companion. We further note that there are also ‘mixed’ examples. So let $f(x) = x^2(x-1)$ and $T_f = T_0 \cup \{\forall x(f(x) = 0 \rightarrow x = 0 \vee x = 1)\}$. Then T_f is axiomatizable by f and the roots 0, 1. We have $f'(0) = 0$, $f'(1) = 1$, and clearly T_f is not a special theory. It is easy to show that the models of T_f are precisely the indecomposable rings without nilpotent elements. But it is shown in PODĘWSKI and REINEKE (to appear) that this theory has no model companion.

Note that $T_x n = T_x 2$ for all $n \geq 2$.

2. General results about algebraically closed models of special theories

In algebraically closed models of special theories one can often prove that $\text{rad}(A) = J(A)$. Hence, in the case where A is a definable ideal and $\text{rad}(A) = J(A)$, $\text{rad}(A)$ is first order definable, too. Cherlin proved for the theory T_0 of all commutative rings that in every finitely generic model R of T_0 $\text{rad}(A) = J(A)$ for every ideal A of R . Clearly, this is not true in every algebraically closed model of T_0 . This immediately follows from:

LEMMA 4. *Let R be an arbitrary algebraically closed model of T_0 , $a \in R$ a non-unit. Then a is an element of $J(\bigcap_{n \geq 1} (a^n))$, i.e. if M is a maximal ideal of R and $\bigcap_{n \geq 1} (a^n) \subset M$, then $a \in M$.*

PROOF: Let a be a non-unit, $a, b \in R$. Let

$$R' := R[x, y]/(1 - xa - y(1 - ba) + axy(1 - ba)).$$

We will show that R' is an extension of R .

Suppose $d = p(xa - xy(a-1) + y(1 - ba) - 1)$ for some $p \in R[x, y]$. Then by comparison of coefficients we have $d = -p^0$ and $p^{x^k}a - p^{x^{k+1}} = 0$. Let n be a natural number such that $p^{x^{k_1}y^{k_2}} = 0$ for all $k_1, k_2 \geq n$. Thus

$p^x a = 0$ and by induction $p^0 a^{n+1} = 0$. Again by comparison of coefficients and induction we get $p^0(1-ba)^{n+1} = 0$. Define $z := p^0(1-ba)^n$. Hence $z(1-ba) = 0$. Therefore $z = z b^{n+1} a^{n+1}$. Since $p^0 a^{n+1} = 0$, it follows that $za^{n+1} = 0$. Thus $z = 0$, i.e. $p^0(1-ba)^n = 0$. By induction it follows that $p^0(1-ba) = 0$. Again we conclude that $p^0 = p^0 b^{n+1} a^{n+1} = 0$. This shows that R' is an extension of R . Since R is algebraically closed in T_0 , there are $x, y \in R$ such that $(1-y(1-ba))(1-ax) = 0$. From this we can conclude that $1-y(1-ba)$ is an element of $\bigcap_{n \geq 1} (a^n)$. This proves our lemma.

Remark. Let R be an algebraically closed model of T_0 and suppose that there is $a \in R$ such that for all n $a^n \notin (a^{n+1})$, i.e. $a \notin \text{rad}(\bigcap_{n \geq 1} (a^n))$. (For example: in every infinitely generic model such an $a \in R$ exists.)

From Lemma 4 we can conclude that $\text{rad}(\bigcap_{n \geq 1} (a^n))$ is a proper subset of $J(\bigcap_{n \geq 1} (a^n))$.

Toffalori showed (see TOFFALORI) that there are infinitely generic models of T_0 with $\text{rad}(A)$ a proper subset of $J(A)$ for some ideal A . With our above arguments this is true for every infinitely generic model of T_0 . In the infinitely generic case we also have a much easier proof of Lemma 4:

PROPOSITION. *Let T be a special theory and let R be an infinitely generic model of T . Then for all non-units $a \in R$: $a \in J(\bigcap_{n \geq 1} (a^n))$. Further, there exists a non-unit $a \in R$ such that $a \notin \text{rad}(\bigcap_{n \geq 1} (a^n))$. (Moreover, there are infinitely many prim ideals which are not maximal.)*

We will omit the proof.

We will now show that in algebraically closed models of special theories some important second-order properties are first-order definable.

LEMMA 5. *Let T be a special theory and let R be an algebraically closed model of T and $a \in R$. Then the following holds:*

- (a) $\text{rad } R = J$,
- (b) $\text{rad}(a) = J(a)$,
- (c) $\text{rad}(\text{Ann}(a)) = J(\text{Ann}(a))$.

PROOF: Clearly, (a) follows from (b).

(b) We will show that: $b \in \text{rad}(a)$ if and only if for all $d, z, x \in R$

$$(1-db)z = 0$$

and $ax = 0$ then $zx = 0$. One direction follows immediately since $\text{rad}(a) \subset J(a)$. For the converse suppose on the contrary that $b \notin \text{rad}(a)$. Let $R' := R[x, y, z]/((1-yb)z, ax)$. Since T is a special theory, clearly R' is a model of T and obviously R' is an extension of R .

We will show that $zx \neq 0$ in R' . Suppose on the contrary that $zx = p(1-yb)z + qax$ for some polynomials $p, q \in R[x, y, z]$. Then by comparison of coefficients we have $1-p^x$ is an element of (a) and

$$p^{xy^{k+1}} - p^{xy^k}b \in (a).$$

Clearly, there is $n \geq 1$ such that $p^{xy^{n+1}} = 0$. Then $p^{xy^n}b \in (a)$ and, by induction, $p^x b^n \in (a)$. Hence $b^{n+1} \in (a)$. This is a contradiction to our assumption that $b \notin \text{rad}(a)$. Thus $zx \neq 0$ in R' . Since R is algebraically closed in T and R' is a model of T extending R , we can choose $x, y, z \in R$ such that $(1-yb)z = 0$, $ax = 0$ and $zx \neq 0$. A contradiction. This proves part (b) of our lemma.

(c) As in part (b) one can show that $b \in \text{rad}(\text{Ann}(a))$ if and only if for all $d \in R$ $\text{Ann}(1-db) \subset \text{Ann}(a)$ if and only if $b \in J(\text{Ann}(a))$. As a corollary to Lemma 5 we have:

THEOREM 6. *Let T be a special theory. Then T has no model companion.*

PROOF: Let R be an algebraically closed model of T . Let $n \geq 1$. Define $R' := R[x, y]/(x^{n+1}y)$. Since T is a special theory, R' is a model of T extending R . Clearly, $(xy)^n \neq 0$ in R' . Since R is algebraically closed in T , there exists $a \in R$ such that $a^{n+1} = 0$ and $a^n \neq 0$.

Let U be a non-principal ultrafilter on ω . Also, let for all $n \in \omega$, $a_n \in R$ such that $a_n^{n+1} = 0$ and $a_n^n \neq 0$. Let $f(n) = a_n$ for all $n \in \omega$. Clearly, f/U is in the Jacobson radical of the ultra-power R^ω/U . Since U is a non-principal ultra-filter, f/U is not nilpotent. Hence R^ω/U is not algebraically closed in T . This proves our theorem.

It follows from Lemma 5(b) that in algebraically closed models of special theories, $\text{rad}(a)$ is a definable ideal. Let R be an arbitrary ring, $a, b \in R$. Then clearly: $b \in \bigcap_{n \geq 1} (a^n)$ if and only if $\forall c (a \in \text{rad}(c) \rightarrow b \in (c))$.

From Lemma 5 we can therefore conclude that in algebraically closed models of special theories, $\bigcap_{n \geq 1} (a^n)$ is a first-order definable ideal.

This will be very important for the next section. With the help of the definability of the infinite intersection $\bigcap_{n \geq 1} (a^n)$ we get some more definability results for algebraically closed models of special theories. For example:

(c) $\bigcap_{n \geq 1} (a^n) \subset (b)$ if and only if $ca^n \in (b)$ for some natural number $n \geq 1$.

If $c = 1$, then we have $\bigcap_{n \geq 1} (a^n) \subset (b)$ if and only if $a \in \text{rad}(b)$: and this is equivalent to $\bigcap_{n \geq 1} (a^n) \subset \text{rad}(b)$.

PROPOSITION 7. *Let T be a special theory and let R be an algebraically closed model of T . Then the following holds:*

- (a) $\bigcap_{n \geq 1} (a^n)$ is first-order definable;
- (b) $\bigcap_{n \geq 1} (a^n) \subset \text{rad}(b)$ if and only if $a \in \text{rad}(b)$ for all $a, b \in R$;
- (c) $\bigcap_{n \geq 1} (a^n) = \bigcap_{n \geq 1} (b^n)$ if and only if $\text{rad}(a) = \text{rad}(b)$;
- (d) $a^n \in (a^{n+1})$ for some $n \geq 1$ if and only if $\bigcap_{n \geq 1} (a^n)$ is a principal ideal.

We will leave the proof to the reader.

3. Generic rings

One of the main theorems for noetherian rings is the Lemma of Nakayama, which we will present in the following easy form:

FACT (LEMMA OF NAKAYAMA). *Suppose that R is a noetherian ring, $a \in R$; then there exists $b \in R$ such that $(1 - ba) \bigcap_{n \geq 1} (a^n) = 0$.*

CHERLIN (1973) proved that for all finitely generic models of T_0 (the theory of all commutative rings), R and for $a \in R$ there is $n \geq 1$ such that $a^n \in (a^{n+1})$. Hence the above Fact is obviously true for every finitely generic model of the theory T_0 of all commutative rings. The following lemma will show the converse for arbitrary algebraically closed models of special theories and will explain the Lemma of Nakayama in algebraically closed models.

LEMMA 8. *Let T be a special theory and let R be an algebraically closed model of T . Let $a \in R$. Then the following are equivalent:*

- (a) There exists $b \in R$ such that $(1 - ba) \bigcap_{n \geq 1} (a^n) = 0$;
- (b) there exists $n \geq 1$ such that $a^n \in (a^{n+1})$.

PROOF: Clearly, (b) implies (a). For the converse suppose that $(1 - ba) \bigcap_{n \geq 1} (a^n) = 0$ for some $b \in R$. Claim that

$$\text{Ann}\left(\bigcap_{n \geq 1} (a^n)\right) = \bigcup_{n \geq 1} \text{Ann}(a^n).$$

Clearly, $\text{Ann}(a^n)$ is a subset of $\text{Ann}(\bigcap_{n \geq 1} (a^n))$ for all $n \geq 1$. So suppose that x is an element of $\text{Ann}(\bigcap_{n \geq 1} (a^n))$. Suppose on the contrary that $xa^n \neq 0$ for all $n \geq 1$. Then $a \notin \text{rad}(\text{Ann}(x))$. From Lemma 5(c) we can conclude that there are $c, y \in R$ such that $(1-ca)y = 0$ and $xy \neq 0$. Hence $y \in \bigcap_{n \geq 1} (a^n)$ and $xy \neq 0$. This is a contradiction to our assumption that x is an element of $\text{Ann}(\bigcap_{n \geq 1} (a^n))$. Hence

$$\text{Ann}(\bigcap_{n \geq 1} (a^n)) = \bigcup_{n \geq 1} \text{Ann}(a^n).$$

Since $1-ba$ is an element of $\text{Ann}(\bigcap_{n \geq 1} (a^n))$, it follows that $1-ba$ is an element of $\text{Ann}(a^n)$ for some $n \geq 1$. This proves the lemma.

The following example shows that for some special theories the Lemma of Nakayama is false for every algebraically closed model.

EXAMPLE. Let T be the theory of indecomposable rings. From Corollary 3 we can conclude that T is a special theory. Let R be an arbitrary algebraically closed model of T . Suppose that for every $a \in R$ there exists $n \geq 1$ such that $a^n \in (a^{n+1})$. Then every prime ideal is maximal. Since R is indecomposable, R would have precisely one maximal ideal. But clearly every algebraically closed model of a special theory has an infinite number of maximal ideals. Thus there exists $a \in R$ such that $a^n \notin (a^{n+1})$ for all $n \geq 1$. From Lemma 8 it follows that $(1-ba) \bigcap_{n \geq 1} (a^n) \neq 0$ for all $b \in R$.

So we have to find a weaker version of the Lemma of Nakayama to get a property which will separate the finitely generic models of arbitrary special theories.

THEOREM 9. *Let T be a special theory and let R be a finitely generic model of T . Then for all $m \geq 1$ and for all $a \in R$ the following holds in R : for all $z_1, \dots, z_m \in \bigcap_{n \geq 1} (a^n)$ there exists $b \in R$ such that $z_1, \dots, z_m \in \text{Ann}(1-ba)$.*

PROOF: Let $a, z_1, \dots, z_m \in R$ and suppose on the contrary that there exists a finite condition $p \subset D(R)$ such that p forces: $p \Vdash "z_1, \dots, z_m \in \bigcap_{n \geq 1} (a^n)"$ and $p \models \forall b \left(\bigvee_{i=1}^m z_i(1-ba) \neq 0 \right)$ where " $z \in \bigcap_{n \geq 1} (a^n)$ " is the first-order formula of Proposition 7.

Since T is universally axiomatizable, there exist a noetherian ring H which is a model of T and an assignment to the constants which satisfies p and $a, z_1, \dots, z_m \in H$.

Since H is a noetherian ring, there exists $b \in H$ such that $(1 - ba) \cap_{n \geq 1} (a^n) = 0$ (in H).

Claim: $z_i \notin \cap_{n \geq 1} (a^n)$ for some $i = 1, \dots, m$. Suppose on the contrary that for all $i = 1, \dots, m$ $z_i \in \cap_{n \geq 1} (a^n)$. Then $(1 - ba)z_i = 0$ for all i . Let

$$q := p \cup \{(1 - ba)z_i = 0; i = 1, \dots, m\}.$$

Clearly, q is a condition extending p and

$$q \Vdash \exists b \left(\bigwedge_{i=1}^m z_i(1 - ba) = 0 \right).$$

This is a contradiction. Hence we can assume without loss of generality that $z_1 \notin \cap_{n \geq 1} (a^n)$ in H . Therefore there is $n \geq 1$ such that $z_1 \notin (a^n)$. Since H is a model of T and T is a special theory, there exists a model H' of T extending H and some $d \in H'$ with $da^n = 0$ and $dz_1 \neq 0$. Let \bar{c} be the constant symbols of p , z_i , a . Let

$$q := p \cup \{da^n = 0, dz_1 \neq 0\}$$

where the new constants do not occur in \bar{c} .

Clearly, H' realize q . Therefore, q is a condition extending p . Obviously, there is no model of T which satisfies q and $\exists x(z_1 = xa^n)$. Hence, by a standard argument, q forces the negation of the above formula, i.e. $q \Vdash \neg \exists x(z_1 = xa^n)$. Thus $q \Vdash \neg \exists x(z_1 = xa^n)$. Clearly,

$$q \Vdash \forall y_1 \forall y_2 \forall y_3 ((1 - y_1 a)y_2 = 0 \wedge y_3 a^n = 0 \rightarrow y_2 y_3 = 0).$$

It follows that

$$\begin{aligned} q \Vdash \exists b & \left(\neg \exists x(z_1 = xb) \wedge \forall y_1 \forall y_2 \forall y_3 \right. \\ & \left. ((1 - y_1 a)y_2 = 0 \wedge by_3 = 0 \rightarrow y_2 y_3 = 0) \right). \end{aligned}$$

This is a contradiction, since p forces the negation of this formula and q is a condition extending p .

This proves our theorem.

As a corollary we get:

COROLLARY 10. *Let T be a special theory and let R be a finitely generic model of T and let $a \in R$. Then the following are equivalent:*

$$(a) z \in \cap_{n \geq 1} (a^n);$$

$$(b) z \in (az).$$

PROOF: $m = 1$ in Theorem 9.

We will now show that in infinitely generic rings of special theories the converse of Corollary 10 is true:

THEOREM 11. *Let T be a special theory and let R be an infinitely generic model of T . Then there are $a, z \in R$ such that $z \in \bigcap_{n \geq 1} (a^n)$ and $z \notin (az)$.*

PROOF: For every $k \geq 1$ and $n_1, \dots, n_k \geq 1$, let

$$R_k := R[y, x_1, \dots, x_k]/(x_1 y^{n_1+1}, \dots, x_k y^{n_k+1}).$$

Since T is a special theory and R is a model of T , R_k is a model of T , too. Since R is algebraically closed in T and R_k is a model of T extending R , there exist for all $k \geq 1$ and n_1, \dots, n_k some $a, x_1, \dots, x_k \in R$ such that $x_i a^{n_i+1} = 0$ and $x_i a^{n_i} \neq 0$ for every $i = 1, \dots, k$. Hence, by the Compactness Theorem, there is an elementary extension R' of R and $a \in R'$ such that for all $n \geq 1$ there is $b_n \in R'$ with $b_n a^n \neq 0$ and $b_n a^{n+1} = 0$, i.e. $\text{Ann}(a^n)$ is a proper subset of $\text{Ann}(a^{n+1})$ for all $n \geq 1$. Since R' is an elementary extension of R , R' is a model of T . Let $z_n := b_n a^n$.

Let U be a non-principal ultrafilter on ω . Define $f(n) := z_n$. Since U is a non-principal ultrafilter, it follows that f/U is an element of $\bigcap_{n \geq 1} (a^n)$ in the ultrapower R'^ω/U . Clearly, $f/U \cdot a = 0$ and $f/U \neq 0$ in R'^ω/U .

Let R'' be an infinitely generic extension of the ultrapower R'^ω/U . Then in R'' the following holds:

There are $a, z \in R''$ with $z \in \bigcap_{n \geq 1} (a^n)$ and $z \neq 0$ and $az = 0$.

Since R'' is an algebraically closed model of T and R is an elementary substructure of R'' , it follows from Proposition 7 that the same statement holds in R . This proves our theorem. As a corollary we get now our main theorem:

COROLLARY 12. *Let T be a special theory. Then there exists an A_3 -sentence which holds for all finitely generic models of T and whose negation holds in all infinitely generic models of T .*

PROOF: This follows immediately from Corollary 10, Theorem 11, and Proposition 7.

Remark. We have found a sufficient condition in which there exists a sentence that separates the theory of finitely generic models and infinitely generic models of T . But this condition is not necessary. For example, let T be the theory of rings which are imbeddable in local rings. Then it

follows from PODĘWSKI and REINEKE (1979) that there is a sentence which holds in all finitely generic models of T and whose negation holds in all infinitely generic models of T . But, clearly, there exists a ring R and $a \in R$ which is imbeddable in a local ring such that $R[x]/(ax)$ is not imbeddable in a local ring. Hence T is not a special theory.

References

- BARWISE, J., and A. ROBINSON, 1970, *Completing theories by forcing*, vol. 2
- CHERLIN, G., 1973, *Algebraically closed commutative rings*, the Journal of Symbolic Logic, vol. 3, pp. 493–499
- LIPSCHITZ, and SARACINO, 1973, *The model companion of the theory of commutative rings without nilpotent elements*, Proceedings of the American Mathematical Society, vol. 38, pp. 381–387
- NAGATA, M., 1962, *Local rings*, Interscience Tracts, vol. 13
- PODĘWSKI, K.-P. and J. REINEKE, *Algebraically closed commutative indecomposable rings* (to appear in: Algebra and Universalis)
- PODĘWSKI, K.-P. and J. REINEKE, 1979, *Algebraically closed local rings*, The Journal of Symbolic Logic, vol. 44, No. 1
- TOFFALORI, C., *Alcune osservazioni sugli anelli commutativi esistenzialmente chiusi*

SMALL DEGREES IN ORDINARY RECURSION THEORY

A. N. DEGTEV

Tymen, U.S.S.R.

The 1- and the reducibilities between m- and tt-reducibility are considered in this report. If an r- is such a reducibility, let L_r be the upper semilattice of *recursively enumerable* (r.e.) r-degrees and $\text{Th}(L_r)$ the elementary theory of L_r .

We call an m-degree *undissolvable* if it contains only one 1-degree (and, consequently, consists only of cylinders). YOUNG (1966) noticed that every m-degree is undissolvable or contains an infinite chain of 1-degrees. This result easily follows from the fact that if A is a non-cylinder, then $A \oplus A$ is also a non-cylinder and $A <_1 A \oplus A$. If A is not a cylinder but an r.e. non-recursive set, let $L(A)$ be a partial order of the 1-degrees contained in the m-degree of A . It is shown in DEGTEV (1976a) that the structures of $L(A)$ are of great variety. In particular, $L(A)$ has an infinite antichain (i.e. infinitely many pair-incomparable elements) and two incomparable elements, whose least upper bound is the greatest element of $L(A)$. If A is a simple set, then $L(A)$ is not an upper or lower semilattice and (as Dekker remarked) has no minimal elements. In DEGTEV (1976a) it is also proved:

- (a) *there is an r.e. set A such that $L(A)$ is a dense lattice with least element;*
- (b) *for each n there is an r.e. set A such that $L(A)$ has the least element 0 and exactly n elements a_1, a_2, \dots, a_n such that*
 - (i) $i \neq j \Rightarrow a_i \not\leq a_j \& a_j \not\leq a_i$;
 - (ii) $(\forall i)(\forall a \in L(A))(a \leq a_i \Rightarrow (a = 0 \vee a = a_i))$.

We call the binary relation η on $N = \{0, 1, \dots\}$ *positive* if $\{\langle x, y \rangle : x \eta y\}$ is an r.e. set. Let η be a positive equivalence relation and $A \subseteq N$. We say that A is η -closed if

$$(\forall x)(\forall y)(x \in A \& x \eta y \Rightarrow y \in A).$$

We call an η -closed non-recursive set A *ideal* if the equivalence η is such that there are only two η -closed recursive sets: \emptyset and N , and we call A η -*maximal* if A is r.e. and for all η -closed r.e. sets B , $B \setminus A$ or $N \setminus B$ contains only finitely many equivalence classes with respect to η . It is obvious that an η -maximal set has a minimal m-degree. ERSHOV (1971) showed that the m-degrees of any ideal set are undissolvable and if A has a recursively inseparable r.e. set $B \subseteq N \setminus A$ (or A is simple non-hypersimple), then " $A = \{n : D_n \cap A \neq \emptyset\}$ " is an ideal set (for a suitable η), where D_n is a finite set with canonical index n . In particular (JOCKUSCH, 1969), every r.e. non-recursive T-degree has an r.e. undissolvable m-degree, because every such T-degree has some simple non-hypersimple set (YATES, 1965). Note that there are r.e. sets which are not ideal whose m-degrees are undissolvable (DEGTEV, 1973; DENISOV, 1974). On the other hand, LACHLAN (1972) proved that every r.e. non-recursive T-degree has an η -maximal set. It may be proved (DEGTEV, 1976b) that every such T-degree has an η -maximal set whose m-degree is undissolvable and, consequently, has a minimal r.e. 1-degree. An example of a r.e. set which is not η -maximal (for all η) but has a minimal m-degree is also constructed in DEGTEV (1976b).

Let $r-$ be a reducibility between m- and tt-reducibility. The main ones of these reducibilities are: m-, bc-, bd-, btt-, c-, d- and tt-. Recall that

$$A \leqslant_c B \Leftrightarrow (\exists \varphi t.r.f.) (\forall x)(x \in A \Leftrightarrow D_{\varphi(x)} \subseteq B),$$

$$A \leqslant_d B \Leftrightarrow (\exists \varphi t.r.f.) (\forall x)(x \in A \Leftrightarrow D_{\varphi(x)} \cap B \neq \emptyset).$$

The bc- and bd- (bounded c- and d-) reducibilities are defined in the natural way. It is known (DEGTEV, 1973) that L_r is not a lattice, has minimal elements and an $a \neq 0$ such that for any $b \leqslant a$, b is not minimal. For L_m these results were established in ERSHOV (1969). The simple proofs may be found in DEGTEV (1972a). The following proposition is also true for all L_r (DEGTEV, 1976b):

$$(\forall a)(\forall b)(a \neq 1 \& a \not\leqslant b \Rightarrow (\exists c)(a \not\leqslant c \& c \not\leqslant a \& b < c)).$$

For $r- = m-$ LACHLAN (1966) proved that

$$A \oplus B \text{ r-complete} \Rightarrow A \text{ r-complete} \vee B \text{ r-complete}.$$

This result holds for $r- \in \{\text{bc-}, \text{c-}\}$, but not for $r- \in \{\text{bd-}, \text{d-}, \text{btt-}, \text{tt-}\}$. Hence $\text{Th}(L_r) \neq \text{Th}(L_R)$ where $r- \in \{m-, \text{bc-}, \text{c-}\}$ and $R \in \{\text{bd-}, \text{d-}, \text{btt-}, \text{tt-}\}$. MARCHENKOV (1976, 1977) proved that $\text{Th}(L_r) \neq \text{Th}(L_R)$ where $r- \in \{\text{bd-}, \text{btt-}\}$ and $R \in \{\text{d-}, \text{tt-}\}$. The author has received the new theorems:

$\text{Th}(L_m) \neq \text{Th}(L_r)$ where $r \in \{\text{bc-}, \text{c-}\}$, and $\text{Th}(L_r) \neq \text{Th}(L_R)$ where $r \in \{\text{btt-}, \text{tt-}\}$ and $R \in \{\text{bd-}, \text{d-}\}$. The following question arises: $\text{Th}(L_{bc}) = \text{Th}(L_c)$?

It is known that every r.e. non-recursive T-degree contains an infinite antichain of r.e. tt-degrees (DEGTEV, 1972b), every r.e. non-recursive tt-degree contains an undissolvable m-degree and every r.e. non-recursive btt-degree has no a maximal m-degree of all the m-degrees which it contains (КОВЗЕВ, 1975). JOCKUSCH (1969) showed that every non-recursive tt-degree has an infinite chain of m-degrees. The author proved (DEGTEV, 1978) the following:

- (i) *every non-recursive tt-degree contains at least two btt-degrees;*
- (ii) *if a T-degree a is such that $a' \geq 0'$, then it has no minimal r.e. tt-degrees.*

The author's latest result in this trend is: *there exists a non-recursive tt-degree which has no undissolvable m-degree*. Is there a btt-degree, which contains only one m-degree?

The proofs of these results may also be found in DEGTEV (1979) and in the author's paper *Some results on upper semilattices and m-degrees* (forthcoming).

References

- DEGTEV (Дёгтев А. Н.), 1972а, *Об m-степенях простых множеств*, Алгебра и логика, т. 11, № 2, стр. 130–139.
- DEGTEV (Дёгтев, А. Н.), 1972 б, *Наследственные множества и табличная сводимость*, Алгебра и логика, т. 11, № 3, стр. 257–268.
- DEGTEV (Дёгтев А. Н.), 1973, *О tt и m-степенях*, Алгебра и логика, т. 12, № 2, стр. 143–161.
- DEGTEV (Дёгтев А. Н.), 1976а, *О частично упорядоченных множествах 1-степеней, содержащихся в р. н. m-степенях*, Алгебра и логика, т. 15, № 3, стр. 249–266.
- DEGTEV (Дёгтев А. Н.), 1976 б, *О минимальных 1-степенях и табличной сводимости*, Сиб. мат. журнал, т. XVII, № 5, стр. 1014–1022
- DEGTEV (Дёгтев А. Н.), 1978, *Три теоремы о tt-степенях*, Алгебра и логика, т. 17, № 3, стр. 270–281
- DEGTEV (Дёгтев А. Н.), 1979, *О свободимостях табличного типа в теории алгоритмов*, Успехи мат. наук, т. 34, № 3, стр. 137–168
- DENISOV (Денисов С. Д.), 1974, *Три теоремы об элементарных теориях и tt-сводимости*, Алгебра и логика, т. 13, № 1, стр. 5–8
- ERSHOV (Ершов Ю. Л.), 1969, *Гипергиперпростые m-степени*, Алгебра и логика, т. 5, стр. 523–552.
- ERSHOV (Ершов Ю. Л.), 1971, *Позитивные эквивалентности*, Алгебра и логика т. 10, № 6, стр. 620–650.

- JOCKUSCH, C., 1969, *Relationships between reducibilities*, Transactions of the American Mathematical Society, vol. 142, 1, pp. 229–237
- KOBZEV (Кобзев, Г. Н.), 1975, бтт-сводимость, диссертация, Новосибирск, стр. 3–69.
- LACHLAN, A. H., 1966, *A note on universal sets*, The Journal of Symbolic Logic, vol. 31, 4, pp. 573–574
- LACHLAN, A. H., 1972, *Two theorems on many-one degrees of r. e. sets*, Algebra and Logic, vol. 11, 2, pp. 216–229
- MARCHENKOV, S. S. (Марченков, С. С.), 1976, *К сравнению верхних полурешеток р. н. табличных степеней и т-степеней*, Мат. заметки, т. 20, № 1, стр. 19–25
- MARCHENKOV, S. S. (Марченков, С. С.), 1977, *О р. н. минимальных бтт-степенях*, Мат. сборник, т. 103, № 4, стр. 550–562.
- YATES, C. E. M., 1965, *Three theorems on the degrees of r. e. sets*, Duke Mathematical Journal, vol. 32.3, pp. 461–468
- YOUNG, P. P., 1966, *Linear ordering under one-one reducibility*, The Journal of Symbolic Logic, vol. 31.1, pp. 70–85

NON-OBTAINABLE CONTINUOUS FUNCTIONALS*

DAG NORMANN

Oslo University, Oslo, Norway

Abstract. For each $k \geq 3$ we construct a continuous functional Δ of type $k+1$ with a recursive associate such that Δ is not Kleene-computable in any continuous functional of type $\leq k$.

1. Introduction

The countable or continuous functionals were first defined independently by KLEENE (1959b) and KREISEL (1959). Kleene's countable functionals are a sub-class of the total functionals while Kreisel's continuous functionals are equivalence-classes of functions $f: N \rightarrow N$. In this paper we will regard the countable functionals as a type-structure $\langle Ct(k) \rangle_{k \in \omega}$ where each $\psi \in Ct(k+1)$ is a total map $\psi: Ct(k) \rightarrow \omega$. This is equivalent to Kreisel's definition and it was also used in e.g. BERGSTRA (1976) and GANDY & HYLAND (1977).

We will work with a fixed $k \geq 3$. We let n, m, k, i, j , etc., denote natural numbers, $f, g, h, \alpha, \beta, \gamma$ will denote elements of $Ct(1)$, F will denote an element of $Ct(k-1)$, φ, ψ will denote elements of $Ct(k)$ and Δ will denote an element of $Ct(k+1)$.

We let $\sigma, \tau, \pi, \delta$ denote finite sequences which we without mentioning will identify with their sequence-numbers. $\sigma(n-1)$ will denote the n th coordinate of σ when $0 < n \leq lh(\sigma)$. We use the standard notation $f(n) = \langle f(0), \dots, f(n-1) \rangle$ and $\bar{\sigma}(n) = \langle \sigma(0), \dots, \sigma(n-1) \rangle$ whenever $n \leq lh(\sigma)$.

* Preprint Series, Matematisk institutt, Universitetet i Oslo, ISBN 82-553-0384-7, Mathematics, No 9, June 1, 1979.

KLEENE (1959b) showed that the class of countable functionals is closed under S1–S9 (KLEENE, 1959a), and he showed that all computable functionals are recursive, i.e. have recursive associates.

Later Tait showed that the converse is not true. The fan-functional Φ is recursive but not computable in any f . Φ is a functional working on two arguments $G \in \text{Ct}(2)$ and f . If

$$C_f = \{g; \forall n g(n) \leq f(n)\},$$

we let

$$\Phi(G, f) = \mu n \quad \forall g_1, g_2 \in C_f (\bar{g}_1(n) = \bar{g}_2(n) \Rightarrow G(g_1) = G(g_2)).$$

Tait never published his result, but sufficient arguments are given in e.g. GANDY & HYLAND (1977), FENSTAD (1980), and NORMANN (1980).

Later Gandy defined a new functional Γ in $\text{Ct}(3)$ as follows:

$$\Gamma(G) = G_0(\lambda n \Gamma(G_{n+1}))$$

where

$$G_n(f) = G(n * f) \quad (* \text{ denotes concatenation}).$$

Gandy showed that Γ is recursive and Hyland showed that Γ is not computable in Φ and any f . The proof is based on some material in BERGSTRA (1976) and can be found in GANDY & HYLAND (1977) and NORMANN (1980).

The following problem still remains open: “Are all continuous functionals computable in an element of $\text{Ct}(3)$?” In this paper we solve this problem by constructing a recursive $A \in \text{Ct}(k+1)$ for all $k \geq 3$ such that A is not computable in any $\varphi \in \text{Ct}(k)$.

2. Conventions and preliminaries

From now on we will use the following notation and conventions:

Let B_σ^n denote the set of functionals in $\text{Ct}(n)$ with an associate extending σ . We will then have

$$B_\delta^1 = \{f; f(\text{lh}(\delta)) = \delta\}.$$

When we use the letters σ and τ we will always assume that $B_\sigma^{k-1} \neq \emptyset$, $B_\tau^{k-1} \neq \emptyset$.

LEMMA 1. a. $B_\sigma^{k-1} \subseteq B_\tau^{k-1} \Leftrightarrow$

$$\forall \delta, s \ (\tau(\delta) = s+1 \Rightarrow \exists \pi B_\delta^{k-2} \subseteq B_\pi^{k-2} \wedge \sigma(\pi) = s+1).$$

- b. If $B_\sigma^{k-1} \subseteq B_{\tau_1}^{k-1} \cup \dots \cup B_{\tau_n}^{k-1}$, then $\exists i \leq n B_\sigma^{k-1} \subseteq B_{\tau_i}^{k-1}$.
- c. If $B_\sigma^{k-1} \not\subseteq B_{\tau_1}^{k-1} \cup \dots \cup B_{\tau_n}^{k-1}$, then there is an extension σ_1 of σ such that $B_{\sigma_1}^{k-1} \cap (B_{\tau_1}^{k-1} \cup \dots \cup B_{\tau_n}^{k-1}) = \emptyset$.

Both this and the next lemma are elementary and we will not prove them here.

LEMMA 2. a. Let $I = \{\sigma; B_\sigma^{k-1} \neq \emptyset\}$. There is a primitive recursive family $\{F_\sigma\}_{\sigma \in I}$ in $Ct(k-1)$ such that

- i. $F_\sigma \in B_\sigma^{k-1}$.
- ii. $F_\sigma = F_\tau \wedge \sigma \neq \tau \Rightarrow B_\sigma^{k-1}$ contains just F_σ which is constant.
- iii. If $\sigma < \tau$ and $B_\tau^{k-1} \not\subseteq B_\sigma^{k-1}$, then $F_\tau \notin B_\sigma^{k-1}$.
- b. There is a primitive recursive dense family $\{\xi_i\}_{i \in N}$ in $Ct(k-2)$ such that the relation $\xi_i \in B_\delta^{k-2}$ is primitive recursive.

For each F we let $h_F(i) = F(\xi_i)$. The following result was essentially first proved in NORMANN (1977). Later S. Dvornickov simplified the proof. His proof is given in NORMANN (a).

LEMMA 3. a. Let $H = \{h_F; F \in Ct(k-1)\}$. Then $H \in \Pi_{k-2}^1 \setminus \Sigma_{k-2}^1$.

b. If A is Π_{k-1}^1 , then there is a primitive recursive R such that

$$\alpha \in A \Leftrightarrow \forall h \in H \ \exists n R(\bar{\alpha}(n), \bar{h}(n), n).$$

DEFINITION. Let $G \in Ct(n)$, $n \geq 2$. We call α a semi-associate for G if $\forall m G \in B_{\alpha(m)}^n$.

In proving the properties of Φ and Γ mentioned above we make use of the following observation:

If $G \in Ct(2)$ then a computation $\{e\}(G)$ depends only on G restricted to a countable set, namely

$$1\text{-sc}(G) = \{f; f \text{ is computable in } G\}.$$

So if α is a semi-associate for G securing all $f \in 1\text{-sc}(G)$, then there is an n such that $\{e\}(G)$ is uniquely determined by $G \in B_{\alpha(n)}^2$. This was proved in GANDY & HYLAND (1977).

Our next lemma gives a higher type version of this observation.

LEMMA 4. Let $\varphi \in Ct(k)$, $\{e\}(\varphi) \simeq s$ by S1-S9. Then there is a Σ_{k-2}^1 -set $A \subseteq H$ such that if $\varphi(F)$ is used in a subcomputation of $\{e\}(\varphi)$, then $h_F \in A$.

PROOF: Let α be an associate for φ . Then the following set C will be $\Sigma_{k-2}^1(\alpha)$:

$C = \{\langle d, \vec{f}, \alpha, \vec{g}, t \rangle; \text{each } f_i, g_j \text{ are associates for functionals } G_i, T_j \text{ of type } \leq k-2 \text{ and } \{d\}(\vec{G}, \varphi, \vec{T}) \simeq t\}$ is a subcomputation of $\{e\}(\varphi)$.

From C it is easy to construct A as we want.

LEMMA 5. Let $\{e\}(\varphi) \simeq s$. Let α be a semi-associate for φ such that whenever $\varphi(F)$ is used in a subcomputation of $\{e\}(\varphi)$, then α secures all associates for F . Then there is an n such that

$$\forall \psi \in B_{\alpha(n)}^k(\{e\}(\psi) \downarrow \Rightarrow \{e\}(\psi) \simeq s).$$

PROOF: The standard proof used when α is an associate will work in this case too.

Remark. Lemmas 4 and 5 may easily be proved for a list $\vec{\varphi}$ of arguments instead of just for φ .

3. The construction

The strategy now is as follows:

1. We construct a recursively compact set K such that
 - i. All $\beta \in K$ are semi-associates for ${}^k 0$.
 - ii. No $\beta \in K$ is an associate.
- iii. If $A \subseteq H$ is Σ_{k-2}^1 , then there is a $\beta \in K$ such that if $h_F \in A$, then β secures all associates for F .
2. For each φ we construct a sequence δ_m^φ uniformly primitive recursive in φ such that $\lim_{m \rightarrow \infty} \delta_m^\varphi$ will be the principal associate for φ .
3. We show that if

$$\Delta_K(\varphi) = \mu n \ \forall m \geq n \ \forall \beta \in K (\beta(m) \neq \delta_m^\varphi),$$

then Δ_K has a recursive associate.

4. If $\forall \varphi (\Delta_K(\varphi) = \{e\}(\varphi, \psi))$ then by Lemma 4 and Lemma 5 there will be a $\beta \in K$ such that $\Delta_K({}^k 0)$ is determined by a finite part $\beta(n)$ of β . We will show that this is as absurd as it seems.

Remark. 1-4 give the main idea behind the construction. In order to carry through the technical arguments we must choose both K and δ_m^φ with some care and define Δ_K in a slightly different way.

From now on let $\Sigma(\alpha, h)$ be the following relation

$$\Sigma(\alpha, h) \Leftrightarrow \exists B \in \Sigma_{k-2}^1(\alpha) \quad (B \subseteq H \wedge h \in B).$$

Then Σ is Π_{k-1}^1 and by Lemma 3.b there is a primitive recursive relation R such that

$$\Sigma(\alpha, h_1) \Leftrightarrow \forall h_2 \in H \ \exists n R(\bar{\alpha}(n), h_1(n), h_2(n), n).$$

For each σ let

$$\sigma_i(\delta) = \begin{cases} (\sigma(\delta)-1)_i + 1 & \text{if } \sigma(\delta) > 0 \\ 0 & \text{if } \sigma(\delta) = 0 \end{cases} \quad i \in \{1, 2\}$$

where $(\)_1$ and $(\)_2$ are the two projection maps of the standard pairing operator \langle , \rangle .

For each σ we let h_σ be the largest sequence such that

$$h_\sigma(i) = s \quad \text{if } \exists \delta (\sigma(\delta) = s+1 \wedge \xi_i \in B_\delta^{k-2}).$$

If B_σ^{k-1} contains more than one element then h_σ is a finite sequence uniformly recursive in σ .

Define

$$P_\alpha(\sigma) = \begin{cases} 1 & \text{if } B_\sigma^{k-1} \text{ contains just one element or if} \\ & \exists n R(\bar{\alpha}(n), h_{\sigma_1}(n), h_{\sigma_2}(n), n), \\ 0 & \text{otherwise.} \end{cases}$$

P_α is uniformly recursive in α and P_α is a semi-associate for ${}^k 0$.

LEMMA 6. a. If $A \subseteq H$ is Σ_{k-2}^1 , then there is an $\alpha \in \{0, 1\}^N$ such that if $h_F \in A$ then α secures all associates for F .

b. P_α is not an associate.

c. If $P_\alpha(\sigma) = 1$ and $B_\sigma^{k-1} \subseteq B_\alpha^{k-1}$, then $P_\alpha(\tau) = 1$.

PROOF: a. Let $\alpha \in \{0, 1\}^N$ be such that A is $\Sigma_{k-2}^1(\alpha)$. Let $B = \{h_1 : h \in A\}$ where $h_1(n) = (h(n))_1$. Then $B \subseteq H$ is $\Sigma_{k-2}^1(\alpha)$ so $\Sigma(\alpha, h_1)$ for all $h \in A$. Let $h_2(n) = (h(n))_2$. Then for $h \in A$

$$\exists n R(\bar{\alpha}(n), h_1(n), h_2(n), n).$$

Let β be an associate for F , $h_F \in A$. Let $h = h_F$. Then

$$h_1 = \lim_{m \rightarrow \infty} h_{(\bar{\beta}(m))_1} \quad \text{and} \quad h_2 = \lim_{m \rightarrow \infty} h_{(\bar{\beta}(m))_2}.$$

It follows that for some m $P_\alpha(\bar{\beta}(m)) = 1$.

b. Let α be given. Let $C = \bigcup \{B \subseteq H : B \text{ is } \Sigma_{k-2}^1(\alpha)\}$. Then $C \subseteq H$ and C is Σ_{k-2}^1 . So there is an $h_1 \in H \setminus C$ and then $\neg \Sigma(\alpha, h)$. Choose $h_2 \in H$ such that $\forall n \neg R(\bar{\alpha}(n), h_1(n), h_2(n), n)$. Let $h_1 = h_{F_1}$, $h_2 = h_{F_2}$ and let β be an associate for $F = \langle F_1, F_2 \rangle$. It is clear that F_1 cannot be constant (since otherwise $\{h_1\} \in \Sigma_{k-2}^1$) so $B_{\beta(n)}^{k-1}$ will always contain more than one element. (If B_σ^{k-1} contains just one element, that element is constant.) It follows that P_α will not secure β .

c. This is trivial from the following monotonicity property:

$$B_\tau^{k-1} \subseteq B_\sigma^{k-1} \Rightarrow h_\sigma < h_\tau$$

which again follows trivially from the definition of h_σ and h_τ . (Use Lemma 1.a.)

This ends the proof of Lemma 6.

Let

$$K_k = \{P_\alpha : \alpha \in \{0, 1\}^N\}.$$

Then K_k is compact and contains only semi-associates for ${}^k 0$ none of which are associates.

We will now show that from such compact sets K we may construct interesting functionals of type $k+1$.

DEFINITION. Let $\varphi \in Ct(k)$. Let δ_m^φ be the sequence of length m defined as follows. For $\sigma < m$ let

$$\delta_m^\varphi(\sigma) = \begin{cases} s+1 & \text{if } \exists \tau < m (\sigma \neq \tau \wedge B_\tau^{k-1} \subseteq B_\sigma^{k-1}) \wedge \\ & \wedge \forall \tau < m (B_\tau^{k-1} \subseteq B_\sigma^{k-1} \Rightarrow \varphi(F_\tau) = s), \\ 0 & \text{otherwise.} \end{cases}$$

LEMMA 7. $\lim_{m \rightarrow \infty} \delta_m^\varphi$ is the principal associate for φ .

The proof is standard.

4. The proof

LEMMA 8. Let K be a compact set of semi-associates for type k functionals such that K contains no associates. Then the functional

$$\Delta_K(\varphi) = \mu n \forall m \geq n \forall \beta \in K \exists \sigma < m (\beta(\sigma) = 0 \wedge \delta_m^\varphi(\sigma) > 0)$$

is well-defined and has an associate recursive in K , i.e. in

$$\{\langle n, \pi_1, \dots, \pi_{k_n} \rangle : \{\pi_1, \dots, \pi_{k_n}\} = \{\beta(n) : \beta \in K\}\}.$$

PROOF: Let α be an associate for φ . It is sufficient to show that $\Delta_K(\varphi)$ is uniformly recursive in α, K .

For each β , if β is a semi-associate and

$$\forall \sigma (\alpha(\sigma) > 0 \Rightarrow \beta(\sigma) > 0),$$

then β is an associate. So

$$\forall \beta \in K \exists \sigma (\beta(\sigma) = 0 \wedge \alpha(\sigma) > 0).$$

Since K is compact, we may choose these σ 's among a finite set $\{\sigma_1, \dots, \sigma_k\}$. Choose m so large that all these sequences have proper extensions $< m$. Then

$$\forall \beta \in K \exists \sigma < m \exists \tau < m (\sigma \neq \tau \wedge B_\tau^{k-1} \subseteq B_\sigma^{k-1} \wedge \beta(\sigma) = 0 \wedge \alpha(\sigma) > 0).$$

Recursively in α, K we may pick m_0 to be the least such m . We then know that

$$\forall m \geq m_0 \forall \beta \in K \exists \sigma < m (\beta(\sigma) = 0 \wedge \delta_m^\varphi(\sigma) > 0).$$

Then $\Delta_K(\varphi)$ is the least $n \leq m_0$ such that

$$\forall m (n \leq m \leq m_0 \Rightarrow \forall \beta \in K \exists \sigma < m (\beta(\sigma) = 0 \wedge \delta_m^\varphi(\sigma) > 0)).$$

We may find this n uniformly recursive in K, m_0 . This shows that Δ_K is recursive in K and ends the proof of Lemma 8.

Let $\Delta_k = \Delta_{K_k}$. Then $\Delta_k \in \text{Ct}(k+1)$ has a recursive associate.

LEMMA 9. Δ_k is not Kleene-computable in any $\psi \in \text{Ct}(k)$.

PROOF: Assume that the lemma is false. Then there is a $\psi \in \text{Ct}(k)$ and an e such that

$$\forall \varphi \in \text{Ct}(k) (\Delta_k(\varphi) = \{e\}(\varphi, \psi)).$$

By Lemma 4 there is a Σ_{k-2}^1 -set $A \subseteq H$ such that whenever ${}^k 0(F)$ is used in a subcomputation of $\{e\}({}^k 0, \psi)$ then $H_F \in A$. By Lemma 6.a there is a $P_a \in K_k$ securing all associates for F whenever $h_F \in A$. By Lemma 5 there is an n such that whenever $\varphi \in B_{P_a(n)}^k$ then

$$\Delta_k(\varphi) = \{e\}(\varphi, \psi) = \{e\}({}^k 0, \psi) = \Delta_k({}^k 0).$$

We defined δ_m^φ for $\varphi \in \text{Ct}(k)$ but we can use the same definition for all φ defined on all F_σ . Let

$$\varphi_0(F_\sigma) = \begin{cases} 0 & \text{if } B_\sigma^{k-1} \text{ contains just one element} \\ & \quad \text{or if } \exists \tau < n (F_\sigma \in B_\tau^{k-1} \wedge P_a(\tau) = 1), \\ \sigma + 1 & \text{otherwise.} \end{cases}$$

By Lemma 2.ii we see that φ_0 is well-defined. Moreover, if

$$\forall \varphi \in B_{P_\alpha(n)}^k (\varphi(F_\sigma) = s)$$

then $\varphi_0(F_\sigma) = s$, so all finite parts of φ_0 may be extended to elements in $B_{P_\alpha(n)}^k$.

CLAIM. a. $\forall m > n \forall \sigma (n \leq \sigma < m \Rightarrow \delta_m^{\varphi_0}(\sigma) \leq P_\alpha(\sigma))$.

b. If $\sigma < n$ and $P_\alpha(\sigma) = 0$, then there is an m_0 such that

$$m \geq m_0 \Rightarrow \delta_m^{\varphi_0}(\sigma) = 0.$$

PROOF: For each m, σ we have that $\delta_m^{\varphi_0}(\sigma)$ is either 0 or $\varphi_0(F_\sigma) + 1$, and $\varphi_0(F_\sigma)$ is either 0 or $\sigma + 1$.

If $\delta_m^{\varphi_0}(\sigma) = \sigma + 2$, then

$$\exists \sigma_1 < m (\sigma_1 \neq \sigma \wedge B_{\sigma_1}^{k-1} \subseteq B_\sigma^{k-1} \wedge \varphi_0(F_\sigma) = \varphi_0(F_{\sigma_1}) = \sigma + 1).$$

But $\varphi_0(F_{\sigma_1}) \neq \sigma + 1$ when $\sigma \neq \sigma_1$ so this is impossible. It follows that $\delta_m^{\varphi_0}(\sigma) \in \{0, 1\}$ for all σ .

a. Assume that $n \leq \sigma < m$ and $\delta_m^{\varphi_0}(\sigma) = 1$. Since $\delta_m^{\varphi_0}(\sigma) > 0 \Rightarrow \delta_m^{\varphi_0}(\sigma) = \varphi_0(F_\sigma) + 1$ for all φ, σ , we must have $\varphi_0(F_\sigma) = 0$. If this is because B_σ^{k-1} contains just one element, we have constructed P_α in such a way that $P_\alpha(\sigma) = 1$.

If B_σ^{k-1} contains more than one element we must have

$$\exists \tau < n (F_\sigma \in B_\tau^{k-1} \wedge P_\alpha(\tau) = 1).$$

Then $\tau < \sigma$ and by Lemma 2.iii we must have $B_\sigma^{k-1} \subseteq B_\tau^{k-1}$. But then by Lemma 6.c $P_\alpha(\sigma) = 1$. So

$$\delta_m^{\varphi_0}(\sigma) = 1 \Rightarrow P_\alpha(\sigma) = 1.$$

b. If $P_\alpha(\sigma) = 0$ then B_σ^{k-1} contains more than one element. If

$$B_\sigma^{k-1} \subseteq \bigcup \{B_\tau^{k-1} : \tau < n \wedge P_\alpha(\tau) = 1\}$$

$$\exists \tau < n (P_\alpha(\tau) = 1 \wedge B_\sigma^{k-1} \subseteq B_\tau^{k-1}).$$

But by Lemma 6.c $P_\alpha(\sigma) = 1$ so this is impossible. So

$$B_\sigma^{k-1} \not\subseteq \bigcup \{B_\tau^{k-1} : \tau < n \wedge P_\alpha(\tau) = 1\}.$$

By Lemma 1.c there are extensions σ_1 and σ_2 of σ such that $\sigma_1 \prec \sigma_2$ and

$$B_{\sigma_1}^{k-1} \cap \bigcup \{B_\tau^{k-1} : \tau < n \wedge P_\alpha(\tau) = 1\} = \emptyset.$$

Then $\varphi_0(F_{\sigma_1}) = \sigma_1 + 1$ and $\varphi_0(F_{\sigma_2}) = \sigma_2 + 1$. For $m > \sigma_2$ we see that $\delta_m^{\varphi_0}(\sigma) = 0$. This ends the proof of the claim.

By the claim we have

$$\exists m_0 > n \quad \forall m \geq m_0 \quad \forall \sigma < m (\delta_m^{\varphi_0}(\sigma) \leq P_a(\sigma)).$$

Choose $m > \max\{\Delta_k(k0), m_0\}$. Let $\varphi \in B_{P_a(n)}^k$ be such that

$$\forall \sigma < m \quad \varphi(F_\sigma) = \varphi_0(F_\sigma).$$

As we remarked after the definition of φ_0 this is possible. Then $\delta_m^\varphi = \delta_m^{\varphi_0}$ and $\forall \sigma < m \quad \delta_m^\varphi(\sigma) \leq P_a(\sigma)$. So $\Delta_k(\varphi) \geq m$. But since $\varphi \in B_{P_a(n)}^k$, we have $\Delta_k(\varphi) = \Delta_k(k0)$. This is a contradiction and the lemma is proved.

We have now showed

THEOREM. *For each $k \geq 2$ there is a recursive functional Δ in $Ct(k+1)$ such that Δ is not computable in any functional in $Ct(k)$.*

PROOF: For $k = 2$ we may use the fan-functional while for $k \geq 3$ we have showed that Δ_k is an example.

References

- BERGSTRA, J., 1976, *Computability and continuity in finite types*, Dissertation (Utrecht)
- FENSTAD, J. E., 1980, *General recursion theory* (Springer Verlag)
- GANDY, R. O., and J. M. E. HYLAND, 1977, *Computable and recursively countable functions of higher type*, in: Logic Colloquium 76, eds. R. O. Gandy and J. M. E. Hyland (North-Holland, Amsterdam)
- KLEENE, S. C., 1959a, 1963, *Recursive functionals and quantifiers of finite types, I; II*; Transactions of the American Mathematical Society, vol. 91, pp. 1–52; vol. 108, pp. 106–142
- KLEENE, S. C., 1959b, *Countable functionals*, in: Constructivity in Mathematics, ed. A. Heyting (North-Holland, Amsterdam), pp. 87–100
- KREISEL, G., 1959, *Interpretation of analysis by means of constructive functionals of finite types*, in: Constructivity in Mathematics, ed. A. Heyting (North-Holland, Amsterdam), pp. 101–128
- NORMANN, D., 1977, *Countable functionals and the analytic hierarchy*, Oslo Preprint, No. 17
- NORMANN, D., 1980, *Recursion on the countable functionals*, Lecture Notes in Mathematics 811 (Springer-Verlag)
- NORMANN, D., a, *Countable functionals and the projective hierarchy* (in preparation)

DYNAMIC LOGIC *

VAUGHAN R. PRATT

Massachusetts Institute of Technology, Cambridge, Mass., U.S.A.

Dynamic logic is a system for reasoning about action. It was originally motivated by the problem of reasoning about computer programs of the iterative kind but has evolved into a more generally applicable system with a mathematically 'clean' structure.

This essay is nominally a tutorial on dynamic logic, but as I feel uncomfortable writing papers devoid of novelty I shall try to combine the tutorial aspect with my current hobby-horse, the algebraization of dynamic logic. Logicians who find the algebraic approach to logic distasteful have my sympathy; algebraic logic can be an uncomfortably abstract way of looking at logic. Nevertheless the power of abstract modes of thought justifies the effort of acclimatization, and the algebraic approach is these days a *sine qua non* of the professional metamathematician. A less algebraic account of dynamic logic may be found in PRATT (to appear).

The following sentence contains an instance of the basic construct of dynamic logic.

By painting a box green I can make that box green.

This proposition composes an action a = 'I paint box green' with a proposition p = 'box is green' to form the proposition ap expressing ' a can bring about p ' or more succinctly ' a enables p '. Dynamic logic deals with propositions, actions, and the 'enables' construction for forming a proposition ap from an action a and a proposition p .

'Enables', or \diamond as we shall call it, has no interesting properties of its own. However, as structure emerges in the propositions and actions, properties of 'enables' also emerge. Suppose the propositions form a Boolean

* This research was supported by NSF Grant No. MCS-7804338.

algebra $\mathcal{B} = (B \vee' 0)$, that is, a complemented distributive lattice, having laws of its own (which may be expressed equationally). There are a number of such equations, which we lump together here as

1. \mathcal{B} is a Boolean algebra.

Then we would expect

$$2a. \quad a0 = 0,$$

that is, it is false that any action can bring about a contradiction.

Furthermore, we would expect

$$2b. \quad a(p \vee q) = ap \vee aq,$$

that is, a can bring about $p \vee q$ just when either a can bring about p or a can bring about q .

If there is exactly one action, this system corresponds to the Kripke system K of modal logic where ap is written $\diamond p$. The reader steeped in modal logic will be aware that the more usual description of this logic is via a Hilbert-style axiomatization of propositional calculus together with the axiom $\Box(p \Rightarrow q) \supset (\Box p \supset \Box q)$ and the inference rule, from p infer $\Box p$, where $\Box p$ is $(\Diamond p)'$, that is, \Box is the dual of \Diamond . It is not hard to show that the axiom has the same force as $\Diamond(p \vee q) \equiv \Diamond p \vee \Diamond q$, which is 2b, while the inference rule captures ‘from $p \equiv 1$ infer $\Box p \equiv 1'$, i.e. $\Box 1 = 1$, or $\Diamond 0 = 0$, which is 2a.

There is an obvious correspondence between algebraic identities and valid formulas: the equation $p = q$ translates to the formula $p \equiv q$ while the formula p translates to the equation $p = 1$. There is also a correspondence that is not too difficult to work out between the use of the inference rule *Modus Ponens* and the usual rules for manipulating algebraic identities taught in high school. Given that these easy connections exist, it suffices to choose one approach in the sequel; we settle for the algebraic one as promised.

To recapitulate, we have thus far supposed that the propositions form a Boolean algebra $\mathcal{B} = (B \vee' 0)$, the actions are undistinguished, merely forming a set R , and \Diamond satisfies $a0 = 0$ and $a(p \wedge q) = ap \wedge aq$. We shall call the two-sorted algebra $\mathcal{M} = (\mathcal{B} R \Diamond)$ a *modal algebra*. (Usually heterogeneous algebras such as this are presented as $(BR \vee' 0 \Diamond)$, that is, carriers B and R first and then operations. We have simply collected $B \vee'$ and 0 together and called it \mathcal{B} in presenting \mathcal{M} .)

A class of algebras defined by a system of equations, e.g. groups, rings, Boolean algebras, modal algebras, is called a *variety*. Varieties are closed under homomorphisms, subalgebras, and direct products. This is because these operations preserve equational identities.

A famous and surprising theorem due to M. Stone says that every Boolean algebra is isomorphic to a *field* of sets, a set of sets that is closed under union and complement. (Equivalently, every Boolean algebra is a sub-algebra of a direct product of two-element Boolean algebras; each element of such a direct product is a ‘bit vector’ indexed by set elements, forming the characteristic function of a subset of the index set.) Thus Boolean algebra can be used as the logic of sets in the weak sense that while \cap and \cup are permitted, \in is not. As such it means that this weak theory of sets coincides with the theory of Boolean algebras, not merely the equational theory but the first-order theory.

Just as Boolean algebras are naturally interpreted as algebras of sets, so are modal algebras naturally interpreted as Kripke structures. Given a set W (to be thought of as possible worlds), a *Kripke structure* on W is a set of subsets of W closed under union and complement and including the empty set, and a set of subsets of W^2 , that is, binary relations on W . Modal logicians tend to be fairly parsimonious about the number of binary relations they will admit in a Kripke structure, feeling quite daring if two such relations can coexist. Nevertheless, having an entire set of binary relations is well-motivated for a theory of actions.

D. KOZEN (1979) has shown that every modal algebra is isomorphic to a Kripke structure. We find this result as remarkable as Stone’s, as it is tantamount to saying that every finitely additive function on a Boolean algebra mimics the behavior of some completely additive function. Hence the first-order theory of Kripke structures, permitting quantification over both propositions and actions, coincides with that of modal algebras. This strengthens Kripke’s completeness result for the modal system K , that every modal formula valid for Kripke structures is a theorem of K . In our algebraic formulation Kripke’s theorem amounts to the coincidence of the equational theories of Kripke structures and modal algebras. (That Kripke’s theorem generalizes from one relation to a set of relations is a triviality.)

We should explain here that there exists for both equational theories and for first order theories a universal completeness theorem. The theorem is that every equation identically true of (resp. every first order formula valid in) all algebras satisfying a given set of equational (resp. first order)

axioms can be proved from those axioms with no more than high school algebra (resp. the classical first-order system). To be precise about the equational case, $x = x$ is the only logical axiom, and the rules are: $x = y \vdash y = x$, $x = y, y = z \vdash x = z$, $x_1 = y_1, \dots, x_n = y_n \vdash f(x_1, \dots, x_n) = f(y_1, \dots, y_n)$, and $x = y \vdash S(x) = S(y)$ where $S(x)$ substitutes terms for variables in x . Thus whenever we can show that the theory (whether equational or first order) of some class C of algebras coincides with the theory of a class of algebras defined by a set of equations or by a set of first-order formulae, then we can immediately produce a complete axiomatization of C . Moreover, if the axioms are equations, then quantifiers may be dispensed with entirely in the proof system and proofs can be carried out purely equationally.

The interest of this for modal algebra is that we automatically get a complete axiom system when we define a class of algebras equationally. That is, the set of equations holding identically for modal algebra can all be proved from the axioms for modal algebra, using high school algebra as defined above. But since this theory is that of Kripke structures, by Kripke's original proof or by Kozen's isomorphism result, we therefore also have a complete axiom system for the equational theory of Kripke structures. By the same token we also have a complete axiom system for the first order theory of Kripke structure, by Kozen's theorem.

If we enlarge yet further our understanding of propositions, say by contemplating particular propositions such as 'the box is green' or ' $x = 5 + y$ ', although we may learn new laws such as $x + y = y + x$ we do not learn any new laws involving \diamond . The problem is that we have not yet encountered any actions whose properties interact through \diamond with such propositions.

To begin with we propose some functions on actions. Having a *choice* of actions a and b can itself be considered a compound action which we shall denote by $a \cup b$. If the choice of a or b enables p , it surely amounts to either a enabling p or b enabling p . Recalling that ap means that a can bring about p , we have

$$3. (a \cup b)p = ap \vee bp.$$

The *sequence* ab consisting of action a followed by action b is another compound action clearly captured by

$$4. (ab)p = a(bp).$$

It is not unreasonable to think of ap as meaning the application of a to p where a is viewed as a function on the Boolean algebra. To be precise we should call them *quasifunctions*, because in general they fail to satisfy extensionality, $\forall p(ap = bp) \rightarrow a = b$. Axioms 3 and 4 can be thought of as defining functionals, or rather quasifunctionals, namely those of pointwise disjunction and composition, respectively. Let us define an equivalence relation \equiv on R such that $a \equiv b$ just when $ap = bp$ for all p in B . Then it can easily be seen that \equiv is a congruence relation on R as far as \cup and ; are concerned.

The *iteration* a^* is the compound action which intuitively consists of a performed an arbitrary number of times. More precisely, a^* can bring about anything a can ($ap \leq a^*p$); it can preserve the *status quo* (can act as the identity action, i.e. is *reflexive*, $p \leq a^*p$); it can achieve nothing by being performed twice that it could not achieve by being performed once (i.e. is *transitive*, $a^*a^*p \leq a^*p$), and yet it is the minimal action satisfying these conditions, that is, it is the reflexive transitive *closure* of a .

While we were able to state reflexivity and transitivity equationally (recall that $p \leq q$ abbreviates the equation $p \vee q \leq q$), it is less clear how we might state closure equationally. In fact this looks like the sort of thing one ought to be able to prove impossible. Surely this is a situation calling for a principle of induction, where even first-order logic ought to prove inadequate. (It is very frustrating not being in a position to bet with the reader.)

Well, here are the equations that do the trick.

$$5a, b. \quad p \vee aa^*p \leq a^*p \leq p \vee a^*(p' \wedge ap).$$

Let us prove forthwith that a^* is the reflexive transitive closure of a .

Let $a!p = \{q | p \vee aq \leq q\}$. Let $\min S$ be the least element of the partially ordered set S when it exists, and undefined otherwise. (This is in contrast to $\wedge S$, the meet of S , which may exist but not be in S .)

We propose the following alternative to 5a, b.

$$5'. \quad a^*p = \min(a!p).$$

Axiom 5' is not an acceptable equational identity for the purpose of defining a variety, because of its use of \min and $!$. However, it provides an excellent metamathematical characterization of $*$, as the following lemma shows.

LEMMA 1. 5a, b and 5' are interchangeable as axioms.

PROOF: (\Rightarrow). Assume 5a, b. 5a asserts that $a^*p \in a!p$. Now consider arbitrary $q \in a!p$. We show that $a^*p \leq q$. We have:

$$\begin{aligned}
 p &\leq q & (p \vee aq \leq q) \\
 \text{Hence } a^*p &\leq a^*q & (2b - \text{expand definition of } \leq) \\
 &\leq q \vee a^*(q' \wedge aq) & (5b) \\
 &= q \vee a^*0 & (p \vee aq \leq q) \\
 &= q. & (2a, 1)
 \end{aligned}$$

(\Leftarrow). Assume $a^*p = \min(a!p)$. Then $a^*p \in a!p$, so 5a holds. For 5b it suffices to show that $p \vee a^*(p' \wedge ap) \in a!p$ since $a^*p \leq q$ for any $q \in a!p$.

$$\begin{aligned}
 p \vee a(p \vee a^*(p' \wedge ap)) &= p \vee (p' \wedge a(p \vee a^*(p' \wedge ap))) & (1) \\
 &= p \vee (p' \wedge (ap \vee aa^*(p' \wedge ap))) & (2b) \\
 &\leq p \vee (p' \wedge ap \vee aa^*(p' \wedge ap)) & (1) \\
 &\leq p \vee a^*(p' \wedge ap). & (5' \rightarrow 5a) \blacksquare
 \end{aligned}$$

As a corollary we infer that \equiv is a congruence relation with respect to $*$ as well, so that we may think of $*$ as a quasifunctional. We now address the question of characterizing which quasifunctional $*$ is. We first need the algebraic notion of a *quasiclosure operator* on R , namely one that is isotonic ($a \leq b$ implies $fa \leq fb$), reflexive ($a \leq fa$), and quasi-idempotent ($ffa \equiv fa$).

LEMMA 2. $*$ is a quasiclosure operator.

PROOF: (Isotonicity) If $a \leq b$ then for all p , $b!p \subseteq a!p$, whence $\min(a!p) \leq \min(b!p)$, thus $a^*p \leq b^*p$, whence $a^* \leq b^*$.

(Reflexivity) $p \leq a^*p$, so $ap \leq aa^*p \leq a^*p$, for all p , whence $a \leq a^*$.

(Quasi-idempotence) $a^*p = \min(a!a^*p) = a^*a^*p$, so $a^*p \in a^{*\dagger}p$. But if $q \in a^{*\dagger}p$, $p \leq q$, so $a^*p \leq a^*q \leq q$, whence $a^*p = \min(a^{*\dagger}p) = a^{**}p$. \blacksquare

We call the quasiclosure system associated with $*$, namely the fixed points of $*$, the *system of asterates*.

There are of course many quasiclosure operators, and merely being one is not a remarkable thing in a variety. So which quasiclosure operator is $*$? We say that the quasifunction a is *reflexive* when $p \leq ap$ for all p , and *transitive* when $aa \leq a$. Thus a is reflexive and transitive when, for all p , $p \vee aap \leq ap$, i.e. $ap \in a!p$, the characterization we use in the next proof.

LEMMA 3. a is an asterate iff a is reflexive and transitive.

PROOF: (\rightarrow) $ap = a^*p \in a!p$.

(\leftarrow) $a^*p = \min(a!p) \leq ap$, and $ap \leq a^*p$, so $a^*p = ap$. \blacksquare

Thus the system of asterates coincides with the set of reflexive transitive quasifunctions, making * reflexive transitive quasiclosure. From all this we infer the following ‘representation theorem’ for dynamic algebras.

THEOREM 4. *Every dynamic algebra is a Boolean algebra \mathcal{B} together with a set of strict finitely additive quasifunctions on \mathcal{B} closed under the quasi-functionals of pointwise disjunction, composition, and reflexive transitive quasiclosure.*

The definition of ‘reflexive transitive closure’ that we have been using is formally correct. However, in the context of binary relations (in Kripke structures) the definition must be used with caution if one’s intuition about the reflexive transitive closure of a binary relation is to be preserved. The standard definition of a^* is the set of all pairs each belonging to a^i for some finite $i \geq 0$.

Consider, for example, a Kripke structure on the set N of natural numbers such that all subsets of N are present as Boolean elements. Let $a_i = \{(u, v) | u \leq v \leq u+i\}$, $a_\infty = \{(u, v) | u, v \in N\}$, and let these be the only actions in the Kripke structure. Then $a_1^* = a_\infty$ since a_∞ is the least (indeed only) reflexive transitive binary relation containing a_1 . Yet our intuition about transitive closure tells us that a_1^* ought to be $\{(u, v) | u \leq v\}$. The problem is that this relation is absent from the Kripke structure and a_∞ is the least relation around to do the job.

In order to make a^* agree with intuition when a is a binary relation it suffices to have *all* subsets of W^2 as actions, thereby ensuring that the standard a^* will be present and so found.

In the case where * has its standard meaning in a Kripke structure we call this dynamic algebra a *regular Kripke structure*. The regular Kripke structures are of interest, because they seem to be the only dynamic algebras needed by computer scientists. Nonstandard notions of * seem to have little relevance to practice.

Accordingly we ask whether the equational theory of regular Kripke structures contains any identities missing from the equational theory of dynamic algebras. This question was first answered in the negative by K. SEGERBERG (1977) and R. PARIKH (1978), who used non-algebraic methods. Generalizing a technique of FISCHER and LADNER (1977), which is itself a generalization of the modal logic technique of filtration, it can be shown (PRATT, 1979a) that every dynamic algebra is a homomorphic image of a subalgebra of a direct product of Kripke structures and Boolean-trivial (one Boolean element) dynamic algebras, which then ensures that

every Boolean identity of Kripke structures (which trivially must be an identity of Boolean-trivial dynamic algebras) will hold for arbitrary dynamic algebras. This is because these three operations on algebras preserve equational identities.

With all of these classes of algebras having the same equational theory it makes sense to focus on properties of this theory. FISCHER and LADNER (1979) showed that membership in the theory is decidable in nondeterministic time c^n for some $c > 1$, and that no deterministic procedure can do better than time d^n in the worst case for some $d > 1$. PRATT (1978) improved the upper bound to deterministic time c^n , which is within a polynomial of the Fischer-Ladner lower bound.

Kripke structures deal with subsets of W and W^2 . It is natural to extend this to W^* and yet further to $W^* \cup W^\omega$. If both Boolean and regular elements are drawn from the power set of $W^* \cup W^\omega$, and \Diamond and ; are interpreted as concatenation, with the other operations each having their obvious interpretation, it can be shown (PRATT, 1979a) that not only are the resulting algebras dynamic but their equational theory coincides with that of Kripke structures, which is surprising to us.

An interesting action that can be obtained as a function of a proposition is the *test* $p?$ which cannot bring about any change, and cannot bring about anything unless p holds. The test may be captured in the axiom

$$6. \quad (p?)q = p \wedge q.$$

In a Kripke structure the test $p?$ would be the set $\{(u, u) | u \in p\}$.

With tests a nice connection may be made between dynamic logic and iterative computer program constructs. We may represent *if p then a else b* as $p?a \cup p'?b$, and *while p do a* as $(p?a)^*p'?$.

Thus far we have perceived all structure in the propositions and actions in terms of functions on these domains. We now consider particular atomic propositions such as ' $x = 5$ ' and 'John is home,' and particular atomic actions such as 'I changed x to 5' and 'John went home.' This represents a shift to first-order dynamic logic in the sense that our models now include individuals forming some domain such as the integers.

The monograph of D. HAREL (1979) on first order dynamic logic provides far more insight into this subject than we can hope to do justice to here. We should also point out the work of SALWICKI (1970) and his school of algorithmic logic (BANACHOWSKI *et al.*, 1977), which has been particularly productive. Algorithmic logic differs from dynamic logic only to the extent that in place of the regular connectives \cup and $*$ of dynamic logic it uses

the constructs *if-then-else* and *while-do*, which are more familiar to programmers, and, at least until recently, its semantics have been that of deterministic programs. CONSTABLE (1977) has also developed a considerable amount of theory for a logic that is essentially identical to algorithmic logic.

The only kind of action we shall contemplate here for first-order dynamic logic is the *assignment*, which brings about a change in the value of a variable. We mention random assignment, $x := ?$, at the end of this paper. Here we describe *specific* assignment, $x := e$ where x is a variable and e any term. This action changes the value of x from what it was to what e was. Thus $x := 1$ changes x to 1, while $x := x + 1$ changes x to what $x + 1$ was.

The appropriate axiom for $x := e$ is

$$7. \langle x := e \rangle p = p_e^x$$

where p_e^x means informally p with all free occurrences of x replaced by e . Thus $\langle x := x + 1 \rangle x = 3 = x + 1 = 3$.

Dynamic logic as a part of logic

A stereotypical approach to logic begins with propositional calculus and then generalizes to first-order logic. By way of showing where dynamic logic might fit into a development of logical concepts I will describe a more leisurely progression.

The mathematics of propositional calculus is that of Boolean algebras, of first-order logic polyadic algebras (HALMOS, 1962), or cylindric algebras (HENKIN, MONK, TARSKI, 1971) if equality is included. A Boolean algebra is a complemented distributive lattice. A polyadic algebra is a Boolean algebra having, in addition to the usual Boolean functions $\wedge, \vee, ',$ various unary functions called quantifiers each of which satisfies $\exists 0 = 0,$ $\exists(p \vee q) = \exists p \vee \exists q,$ $\exists \exists p = \exists p,$ $\exists(\exists p)' = \exists p'.$ In addition one needs further properties that recognize distinctions between quantifiers.

An algebraic treatment tends to savor this progression, by drawing it out with various intermediate steps and digressions. There is no need to leap headlong into Boolean algebra—one can start out more slowly with quasi-orders, then partial orders, then lattices, then distributive lattices, digress to modular lattices, then visit relatively pseudo-complemented lattices to amuse the intuitionists, and at last arrive at complemented distributive lattices, that is, Boolean algebras. My own preference as a logician

is simply to start with Boolean algebras and move slowly from there, as there is at least as much scope for dilly-dallying on the high side of Boolean algebras as represented by modal logic and beyond than on the low side as represented by lattice theory.

The modalities of modal logic and the quantifiers of first-order logic can be both be considered functions on Boolean algebras. The resemblance goes beyond that however, as can be seen from the properties $f0 = 0$ and $f(x \vee y) = fx \vee fy$ which both enjoy, where in modal logic f is \Diamond while in first-order logic f is \exists . For modal logic, at least for the basic system K , this is all there is to \Diamond . For the modal logic system T we also have $x \leqslant fx$ (where $x \leqslant y$ means $x \vee y = y$; think of it as meaning that x implies y), for $S4$ we have $fx \leqslant fx$, and for $S5$ we have $f(fx)' \leqslant x'$. Quantifiers enjoy all these properties, so first-order logic is an $S5$ modal logic. However, quantifiers satisfy other properties besides these; moreover modal logic tends to have few, usually one, modal function while first-order logic usually has infinitely many quantifiers.

Dynamic logic fits into this picture by virtue of being a system K modal logic with many modalities that comes equipped with functions on the modalities, namely \cup ; and $*$. We might also consider adding φ (satisfying $\varphi x = 0$), and $\bar{}$ or *converse*, satysfying $a^-(ap)' \leqslant p'$ and $a(a^-p)' \leqslant p'$. With these functions a single dynamic logic modality can be made to act like a T , $S4$, or $S5$ modality, as pointed out by FISCHER and LADNER (1977). Thus $A \cup \varphi^*$ supplies a single modality for a T logic, since φ^* is just the multiplicative identity and so $A \cup \varphi^*$ is the reflexive closure of A , where A alone would just be a K modality. Similarly, A^* supplies $S4$ while $(A \cup A^-)^*$ supplies $S5$.

Quantifiers themselves may be considered possible modalities for dynamic logic, in the spirit of thinking of its modalities as actions. The action corresponding to the quantifier $\exists x$ is that of setting the value of variable x to an arbitrarily chosen value. Thus $\exists xP(x)$ says that it is possible to set x to some value to bring about $P(x)$. The appropriate programming notation for this would be $x := ?$, *random assignment*.

References

- BANACHOWSKI, L., A. KRECZMAR, G. MIRKOWSKA, H. RASIOWA, A. SALWICKI, 1977,
An introduction to algorithmic logic; Metamathematical investigations in the theory of programs, in: Mathematical Foundations of Computer Science, eds. A. Mazurkiewicz and J. Pawlak, Banach Center Publications, vol. 2 (Warsaw), pp. 7-99

- CONSTABLE, R. L., 1977, *On the theory of programming logics*, Proceedings of the 9th Annual ACM Symposium on Theory of Computing, Boulder, Col., May, pp. 269-285
- FISCHER, M. J. and R. E. LADNER, 1977, *Propositional modal logic of programs*, Proceedings of the 9th Ann. ACM Symposium on Theory of Computing, Boulder, Col., May, pp. 286-294
- HALMOS, P. R., 1969, *Algebraic logic*, (Chelsea, New York)
- HAREL, D., 1979, *First order dynamic logic*, (Springer-Verlag)
- HENKIN, L., J. D. MONK, and A. TARSKI, 1971, *Cylindric algebra*, (North-Holland, Amsterdam)
- KOZEN, D., 1979, *A representation theorem for models of *-free PDL*, (manuscript, c. May)
- PARIKH, R., 1978, *A completeness result for PDL*, Symposium on Mathematical Foundations of Computer Science, Zakopane, Sept. (Warsaw)
- PRATT, V. R., 1976, *Semantical considerations on Floyd-Hoare logic*, Proceedings of the 17th Ann. IEEE Symposium on Foundations of Computer Science, October, pp. 109-121
- PRATT, V. R., 1978, *A near optimal method for reasoning about action*, MIT/LCS/TM-113, M. I. T., Sept.
- PRATT, V. R., *Applications of modal logic to programming*, Studia Logica (to appear)
- PRATT, V. R., 1979a, *Models of program logics*, submitted to the 20th IEEE Conference on Foundations of Computer Science
- PRATT, V. R., 1979b, *Dynamic algebras: examples and constructions*, Internal Report, M. I. T. Laboratory for Computer Science, May
- SALWICKI, A., 1970, *Formalized algorithmic languages*, Bulletin de l'Académie Polonaise des Sciences, série des sciences mathématiques, astronomiques et physiques, vol. 18, pp. 227-232
- SEGERBERG, K., 1977, *A completeness theorem in the modal logic of programs*, Preliminary report, Notices of the American Mathematical Society, vol. 24, 6, A-552, October

FOUR TEST PROBLEMS IN GENERALIZED RECURSION THEORY *

STEPHEN G. SIMPSON

The Pennsylvania State University, University Park, Pennsylvania, 16802 U.S.A.

By *generalized recursion theory* (abbreviated G.R.T.) I understand any kind of recursion theory in which the domain of individuals is greater than ω . All recursion theory is concerned with computations in which individuals are manipulated by some sort of abstract computing engine. However, the principle of parity (SACKS, 1980) demands that the computation be of approximately the same complexity as the individuals. This means in practice that a particular kind of recursion theory is often determined completely by its domain of individuals. For example, in classical recursion theory (abbreviated C.R.T., SOARE, 1978), the domain of individuals is ω and this fact dictates also the nature of the computations.

G.R.T. is a huge subject. I shall limit myself to only four particular kinds of G.R.T.: α -recursion theory, β -recursion theory, classical descriptive set theory, and recursion theory on an admissible set. Each of these will be discussed from the viewpoint of four carefully selected test problems suggested by C.R.T.

The relations among the four kinds of G.R.T. as well as C.R.T. are summarized in diagram A. The basic notions of recursion theory (RE set, finite set, relative recursiveness) as they apply to each kind of G.R.T. will now be discussed.

(1) *α -recursion theory.* Here and in the literature, α is an admissible ordinal, i.e. an ordinal such that L_α , the α th level of the constructible hierarchy, satisfies Σ_1 replacement. The domain of individuals is α itself,

* Preparation of this paper was partially supported by NSF grant MCS 77-13935. The author is an Alfred P. Sloan Research Fellow.

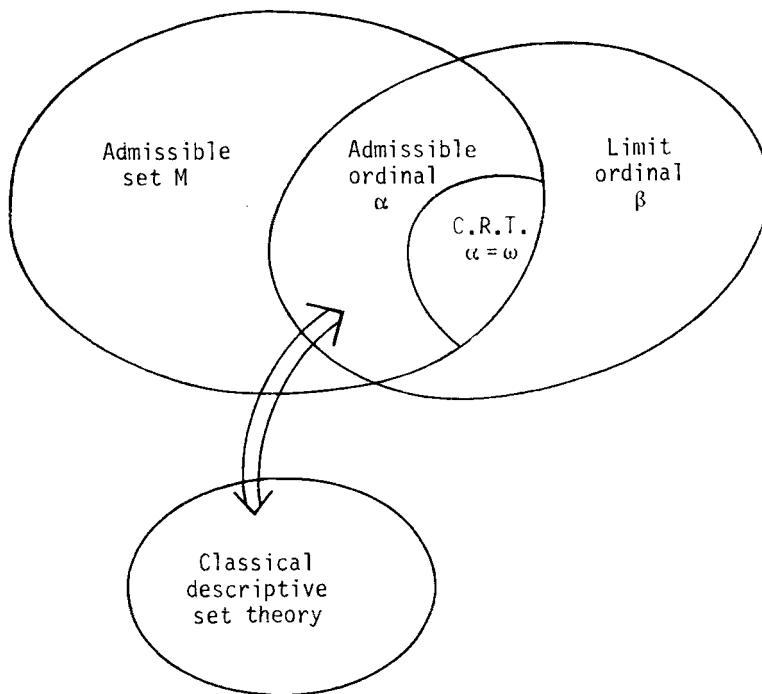


Diagram A

i.e. the set of ordinals less than α . A subset A of α is said to be α -recursively enumerable (abbreviated α -RE) if it is $\Sigma_1(L_\alpha)$, i.e.

$$A = \{x < \alpha : L_\alpha \models \varphi(x, p)\}$$

where φ is a Σ_1 formula and p is a parameter from α . (In C.R.T. the use of parameters is superfluous since each element of ω is definable by Σ_1 formula with no parameters. However, in most kinds of G.R.T. the use of individual parameters is crucial.) As in C.R.T., an α -RE set can also be characterized as the range of an α -recursive function.

A subset of α is said to be α -finite if it is an element of L_α . The notion of relative α -recursiveness (i.e. relation $A \leq_s B$ between subsets A and B of α) can then be defined in a way which is by now fairly uncontroversial. Further information on this highly developed branch of G.R.T. may be found in the Oslo G.R.T. volumes I and II (FENSTAD and HINMAN, 1974; FENSTAD, GANDY and SACKS, 1978) and in the survey article by SHORE (1977) in the Handbook of Mathematical Logic. Note that α -recursion theory contains C.R.T. as the special case $\alpha = \omega$.

(2) β -recursion theory. Here and in the literature, β denotes an arbitrary limit ordinal. The domain of individuals is β . A subset A of β is called β -recursively enumerable (abbreviated β -RE) if A is $\Sigma_1(S_\beta)$ where S_β is the β th level of Jensen's version of the constructible hierarchy. (If $\beta = \omega \cdot \beta$ then $S_\beta = L_\beta$.)

In β -recursion theory there are two competing notions of finiteness: (i) A subset of β is called β -finite if it is an element of S_β . This is the concept used by FRIEDMAN and SACKS (1977) and it leads to a notion of relative β -recursiveness denoted by \leq_β . (ii) Recently Maass has defined a subset of β to be I -finite if it is β -finite and remains so under β -recursive permutations of β . This leads to a notion of relative recursiveness which is denoted by \leq_I . The results obtained by Maass seem to indicate that these notions are also viable and have the additional advantage of meshing well with the axiomatic scheme of Moschovakis and FENSTAD (1980).

(3) Classical descriptive set theory. By this we mean the theory of Borel and analytic sets as developed for instance in KURATOWSKI's book (1966). This subject is not usually regarded as a branch of recursion theory, but we can so regard it under the following identification of basic notions. The domain of individuals is the real line R (or any uncountable complete separable metric space). The role of recursively enumerable sets is played by the coanalytic, i.e. boldface π^1_1 , subsets of R . There is a strong link to admissibility here since $A \subseteq R$ is coanalytic if and only if

$$A = \{x \in R : L_{\omega_1^{x,p}}(x, p) \models \varphi(x, p)\}$$

where φ is a Σ_1 formula, p is a real parameter, and $\omega_1^{x,p}$ is the least ordinal $> \omega$ which is admissible in the pair x, p .

The classical descriptive set theorists obtained a number of structural results on coanalytic sets. Perhaps the most famous of these is Souslin's theorem: *if both A and its complement are coanalytic, then A is Borel.* Thus we are dealing with a kind of recursion theory in which the recursive sets 'are' the Borel sets.

An important conceptual gap in classical descriptive set theory was the lack of a notion corresponding to relative recursiveness. This gap has been filled by Kleene in his work on recursion in higher types. Given $A, B \subseteq R$, we say that A is *Kleene reducible* to B (abbreviated $A \leq_K B$)

$$\chi_A = \lambda x \{e\}(x, p, B, {}^2E)$$

Table B

| | C. R. T. | α -recursion theory | β -recursion theory | | Classical descriptive set theory | M -recursion theory |
|------------------------|-----------------------------|---------------------------------|---------------------------|---------------------|----------------------------------|-----------------------|
| Domain of individuals | ω | admissible ordinal α | limit ordinal β | | R | admissible set M |
| RE set | range of recursive function | $\Sigma_1(L_\alpha)$ | $\Sigma_1(S_\beta)$ | | coanalytic | $\Sigma_1(M)$ |
| Inseparable RE sets | trivially yes | trivially yes | trivially yes | | classical result | Abramson 1979 (d) |
| Finite set | finite | $\in L_\alpha$ | $\in S_\beta$ (Sacks) | I -finite (Maass) | countable set | $\in M$ |
| Maximal RE set | Friedberg 1958 | Lerman 1973 (a) | ? | ? | no (f) | ? |
| Relative recursiveness | \leq_T | \leq_α (Sacks) | \leq_β (Sacks) | \leq_I (Maas) | \leq_K (Kleene) | \leq_M |
| Jump operator | $e \in W_e^A$ | α -jump (Simpson, Shore) | β -jump | I -jump | Super-jump (Gandy) | M -jump |
| Post's Problem | Friedberg-Muchnik 1957 | Sacks-Simpson 1972 | Friedman (b) | Maass (c) | Yes and no (g) | Yes and no (e) |
| Density theorem | Sacks 1964 | Shore 1974 | ? | ? | yes if $V = L$ (h) | ? |

where p is a real parameter, the bracket notation $\{e\}$ refers to Kleene's schemata S1–S9, and 2E is the type 2 number quantifier. This definition may be reformulated in terms of admissibility as follows:

$$A = \{x : L_{\omega_1^{x,p}, B}(x, p, B) \models \varphi_1(x, p)\}$$

$$= \{x : L_{\omega_1^{x,p}, B}(x, p, B) \models \sim \varphi_0(x, p)\}$$

where φ_0 and φ_1 are Σ_1^B formulas and $L_{\omega x, p, B}(x, p, B)$ is the smallest B -admissible set containing x , p , and ω as elements. A and B are said to have the same *Kleene degree* if $A \leqslant_K B$ and $B \leqslant_K A$.

(4) *Recursion theory on an admissible set.* Let M be an arbitrary admissible set (without urelements). In M -recursion theory, the domain of individuals is M and a subset of M is said to be M -RE if it is $\Sigma_1(M)$. An M -finite set is one which belongs to M . The notion of relative M -recursiveness ($A \leqslant_M B$ for $A, B \subseteq M$) can be defined as for α -recursion theory. Note that α -recursion theory is the special case $M = L_\alpha$.

There are of course many kinds of G.R.T. other than the four listed above. My choices here have been dictated mainly by my incompetence in almost all other areas of G.R.T.

At this point it is perhaps appropriate to respond to the obvious question: Why should we generalize recursion theory? This question has been discussed at length by KREISEL (1971). I think that there are two main answers:

(i) *connections to other parts of logic*, e.g. infinitary logic and set theory. In the case of infinitary logic, the history suggests that the logic arose out of the recursion theory and that some of the principal logical results (e.g. Barwise completeness, compactness, and interpolation) were first suggested by recursion-theoretic considerations. In the case of descriptive set theory, it is natural to use the paradigm of C.R.T. to suggest new notions and problems. There is also the well-known link between α -and β -recursion theory and Jensen's fine structure theory, viz. projecta.

(ii) *generalization for generalization's sake.* The classical paradigm here is the well-developed theory of RE sets and degrees in C.R.T. (SOARE) When one contemplates a theory as beautiful and rich as C.R.T. one experiences an uncontrollable urge to generalize. We need not apologize from this urge. Each kind of G.R.T. is equipped with a set of basic notions analogous to those of C.R.T. (viz. RE set, finiteness, relative recursiveness). Therefore, each theorem of C.R.T. has several appropriate generalizations each of which may or may not be true. Thus we have a rich source of interesting problems.

We now turn to our test problems. These will be stated in the form of four well-known theorems of C.R.T.

(A) *Inseparable RE sets.* This is the well-known theorem that there exist two disjoint RE sets A and B such that there is no recursive set X with

$A \subseteq X$ and $X \cap B = \emptyset$. For example, we may take A and B to be the provable and refutable sentences of Peano arithmetic.

(B) *Maximal RE set.* This is Friedberg's theorem that there exists an RE set whose complement is infinite but cannot be split into two infinite parts by an RE set.

(C) *Post's problem.* This is the classical problem of whether there exists an RE set A such that $0 <_T A <_T 0'$. This was answered affirmatively by Friedberg and Muchnik who actually proved more: There exist RE sets A and B which are incomparable under \leq_T .

(D) *Density theorem.* This is the 1964 theorem of Sacks saying that

$$a < b \Rightarrow \exists c \quad (a < c < b)$$

where a , b , and c range over RE degrees.

These four test theorems (A)–(D) have been chosen because of their central role in C.R.T. The rest of the talk will consist of a review of the work that has been done on generalizing each of the test theorems to each of α -recursion theory, β -recursion theory, classical descriptive set theory, and M -recursion theory. The results are summarized in Table B and the accompanying notes (a)–(h) below.

(a) Lerman has shown that the existence of a maximal α -RE set is equivalent to the assertion that α is effectively countable in a certain precise sense.

(b) FRIEDMAN and SACKS (1977) have shown that for many limit ordinals β there exist β -RE sets which are incomparable under \leq_β . On the other hand, FRIEDMAN (to appear) has shown that for many β this result fails, and so Post's problem has a negative solution. This negative result uses an idea borrowed from Silver's work on the singular cardinal problem in set theory.

(c) Maass has recently shown that Post's problem for \leq_I always has an affirmative solution: There exist β -RE sets which are incomparable under \leq_I .

(d) ABRAMSON (to appear) has recently made a study of inseparable M -RE sets. Such sets exist whenever M is sufficiently well-behaved (e.g. $M = L_\alpha$ or M locally countable and closed under hyperjump). However,

Abramson has also shown that there exists a locally countable admissible set $M = M_\alpha$ of any countable height $\alpha > \omega$ such that inseparable M -RE sets do not exist. This result is remarkable, because it points to the existence of admissible logics and recursion theories which are free of some of the 'pathology' associated with Rosser's theorem and inseparability.

(e) Harrington has shown that there exists an admissible set M such that every M -RE set is either M -recursive or complete. Harrington's proof uses an unpublished, rather complicated, forcing construction. Simpson had observed earlier that the same result follows easily if we assume the (false) axiom of determinacy. On the other hand, STOLtenberg-HANSEN (1979) has shown that the Friedberg-Muchnik theorem holds for sufficiently well-behaved admissible sets M .

(f) Maximal coanalytic sets cannot exist because of the classical result that an uncountable analytic set contains a perfect subset. This suggests that perhaps the lattice of coanalytic sets is decidable.

(g) HARRINGTON (1978) has shown that the set theoretical hypothesis $\forall x \exists x^*$ is equivalent to the negative answer to Post's problem for Kleene degrees. HRBACEK and SIMPSON (1980) have shown that if $V = L[G]$ where G is generic with respect to any notion of forcing $P \in L$, then there exists $2^{+\aleph_0}$ coanalytic sets which are pairwise incomparable under \leq_k . This solution of Post's problem has as a byproduct the solution of a classical problem posed by Kuratowski: There exist $2^{+\aleph_0}$ analytic sets no two of which are Borel isomorphic.

(h) HRBACEK (1980) has shown that if $V = L$, then the Kleene degrees of coanalytic sets are dense. He also has a number of other results of the same kind, all of which tend to support the following delightful conjecture: if $V = L$, then the Kleene degrees of coanalytic sets form a universal homogeneous upper semilattice with 0 and 1. This conjecture is of course the precise analog of Shoenfield's conjecture from C. R. T. and would, if true, completely determine the structure of the Kleene degrees of coanalytic sets, up to isomorphism.

References

- ABRAMSON, F., *Locally countable models of Σ_1 separation*, The Journal of Symbolic Logic (to appear)
- FENSTAD, J. E., 1980, *Generalized recursion theory: an axiomatic approach* (Springer-Verlag)

- FENSTAD, J., and P. HINMAN (eds.), 1974, *Generalized recursion theory* (North-Holland)
- FENSTAD, J., R. GANDY, and G. SACKS (eds.), 1978, *Generalized recursion theory* (North-Holland)
- FRIEDMAN, S., and G. SACKS, 1977, *Inadmissible recursion theory*, The Bulletin of American Mathematical Society, vol. 83, pp. 255–256
- FRIEDMAN, S., *Negative solution to Post's problem*. (to appear)
- HARRINGTON, L., 1978, *Analytic determinacy and $0^\#$* , The Journal of Symbolic Logic, vol. 43, pp. 685–693
- HRBACEK, K., 1978, *On the complexity of analytic sets*, Z. math. Logik u. Grundlagen der Math. 24, pp. 419–425
- HRBACEK, K., and S. G. SIMPSON, 1980, *On Kleene degrees of analytic sets*, in: The Kleene Symposium, pp. 269–274 (North-Holland)
- KREISEL, G., 1971, *Some reasons for generalizing recursion theory*, Logic Colloquium '69, pp. 139–198 (North-Holland)
- KURATOWSKI, K., 1966, *Topology*, vol. I (Academic Press and PWN—Polish Scientific Publishers, New York-London and Warszawa)
- SACKS, G., 1980, *Post's problem, absoluteness, and recursion in finite types*, in: The Kleene Symposium (North-Holland)
- SHORE, R. A., 1977, *α -recursion theory*, in: Handbook of Mathematical Logic, pp. 653–680 (North-Holland)
- SOARE, R. I., 1978, *Recursively enumerable sets and degrees*, Bulletin of the American Mathematical Society, vol. 84, pp. 1149–1181
- STOLTENBERG-HANSEN, V., 1979, *Finite injury arguments in infinite computation theories*, Annals of Mathematical Logic, vol. 16, pp. 57–80

ORDINAL GAMES AND THEIR APPLICATIONS

LEO A. HARRINGTON

Department of Mathematics, University of California, Berkeley, California, 94720, U.S.A.

and

ALEXANDER S. KECHRIS *

Department of Mathematics, California Institute of Technology, Pasadena, California, 91125 U.S.A.

In this paper we give a survey of certain results on the determinacy of games on ordinals and its various applications. A more comprehensive treatment including more detailed versions, extensions and proofs of the theorems mentioned here, will appear in a forthcoming paper by the authors.

We would like to thank Ramez Sami for many illuminating discussions on the topics discussed here.

1. Ordinal games

Let λ be an ordinal. To each $A \subseteq \lambda^\omega \times \lambda^\omega$ we associate as usual the following infinite, two person, perfect information game $G(A; \lambda)$:

I ξ_0 ξ_1 ... $\vec{\xi}$ Players I, II alternatively play ordinals
II η_0 η_1 ... $\vec{\eta}$ $\xi_0, \eta_0, \xi_1, \eta_1, \dots < \lambda$; I wins iff $(\vec{\xi}, \vec{\eta}) \in A$.

When $\lambda < \Theta =$ first ordinal (other than 0) not the surjective image of ω^ω , there is a norm $\varphi: \omega^\omega \rightarrow \lambda$, which we view of course as a coding system for ordinals $< \lambda$ by reals ($x \in \omega^\omega$ coding $\varphi(x)$). To simplify the notation, we put $\varphi(x) = |x|$, when there is no danger of confusion. Modulo

* Work partially supported by NSF Grant MCS 76-17254 AO1. The author is an A. P. Sloan Foundation Fellow.

such a norm, we can also consider the coded version of $G(A; \lambda)$, which we shall denote by $G^*(A; \omega^\omega)$, instead of the more accurate $G(A^*; \omega^\omega)$, where $A^* = \{(\vec{\alpha}, \vec{\beta}): ((|\alpha_0|, |\alpha_1|, \dots), (|\beta_0|, |\beta_1|, \dots)) \in A\}$. This game is played as follows:

$$\begin{array}{llllll} \text{I} & u_0 & u_1 & \dots & \vec{u} & \text{I, II alternatively play reals } u_0, v_0, u_1, v_1, \dots \\ \text{II} & v_0 & v_1 & \dots & \vec{v} & \dots \in \omega^\omega; \text{ I wins iff } (|\vec{u}|, |\vec{v}|) \in A \text{ (where} \\ & & & & & |\vec{u}| = (|u_0|, |u_1|, \dots)) \end{array}$$

Clearly, by the Axiom of Choice (AC), these two games are equivalent, i.e. I (II) has a winning strategy in $G \Leftrightarrow$ I (II) has a winning strategy in G^* . This is because, by AC, there is a function $F: \lambda \rightarrow \omega^\omega$ which picks for each ordinal $\xi < \lambda$ a code $F(\xi) = w$ of ξ .

As it turns out, when the norm φ is projective (or, in fact, even hyperprojective), the game $G^*(A; \omega^\omega)$ can in turn be simulated by a game $G(H(A^*); \omega) \equiv HG(A; \omega)$ on ω (i.e. I (II) has a winning strategy in $HG(A; \omega) \Rightarrow$ I (II) has a winning strategy in $G^*(A; \omega^\omega)$). Here

$$H: \text{power } ((\omega^\omega)^\omega \times (\omega^\omega)^\omega) \rightarrow \text{power } (\omega^\omega \times \omega^\omega)$$

is an appropriate operation which depends only on the ‘complexity of φ ’ and not φ itself. For example, if φ is a Δ_n^1 -norm, then for each $X \subseteq (\omega^\omega)^\omega \times (\omega^\omega)^\omega$,

$$H(X) = \{(\alpha, \beta): C(\alpha, \beta) \vee X(\vec{f}(\alpha), \vec{g}(\beta)),$$

where $C \in \Delta_{n+1}^1$, $\vec{f} = (f_0, f_1, \dots)$, $\vec{g} = (g_0, g_1, \dots)$ with $f_i(\alpha) = \beta$ and $g_i(\alpha) = \beta$, Δ_{n+1}^1 , and C, \vec{f}, \vec{g} are independent of φ (i.e. the same C, \vec{f}, \vec{g} work for each φ which is a Δ_n^1 -norm). Thus we have, abbreviating by

$$\text{Proj}[A],$$

the smallest class of sets of reals containing A and the projective sets and closed under \wedge , \vee and substitutions by projective functions:

1.1. THEOREM. *Let $\varphi: \omega^\omega \rightarrow \lambda$ be a projective norm. If $A \subseteq \lambda^\omega \times \lambda^\omega$, then the game $G^*(A; \omega^\omega)$ is determined, granting Determinacy ($\text{Proj}[A]$). (Similarly with hyperprojective instead of projective.)*

The first result on the determinacy of ordinal games was established in HARRINGTON (1976), who proved a special case of the above result for certain Δ_3^1 -norms (and $\lambda = u_\omega =$ the ω th uniform indiscernible). Our method here is an improvement of his basic simulation technique.

As a corollary we have,

1.2. COROLLARY. Let $\kappa = \kappa^R$ be the first non-hyperprojective ordinal. Then

(i) AD \Rightarrow For all $A \subseteq \kappa^\omega \times \kappa^\omega$, $G^*(A; \kappa)$ is determined.

(ii) If all games on ω which are definable from a countable sequence of ordinals are determined and AC holds, then for all such $A \subseteq \kappa^\omega \times \kappa^\omega$, $G(A; \kappa)$ is determined.

Remark: Strictly speaking, 1.1 covers only all $\lambda < \kappa^R$; κ^R itself can be handled by the same method.

Recall here that with AD, κ^R is a very large uncountable ordinal, e.g. it is (weakly) Mahlo.

At this stage the authors do not know how to extend these results to higher $\kappa < \Theta$ (except from some obvious extensions, as e.g. in 1.2 (i) which clearly holds also for all κ 's of the same cardinality as κ^R). It is not known also if there is a definable from a countable sequence of ordinals $A \subseteq \kappa^\omega \times \kappa^\omega$, for any κ whatsoever (not necessarily $< \Theta$), with $G(A; \kappa)$ not determined. Perhaps the assertion of the determinacy of all such games for higher and higher κ 's has some interesting consequences.

An interesting and very useful byproduct of 1.1 is an estimate for the winning strategies in the games G^* . We put down only the projective case, which is mostly used in the applications.

1.3. THEOREM. For each Δ_n^1 -norm $\varphi: \omega^\omega \rightarrow \lambda$ and each $A \subseteq \lambda^\omega \times \lambda^\omega$ there is a Δ_n^1 winning strategy (for whoever player wins) in the game $G^*(A; \omega^\omega)$, assuming Determinacy (Proj[A]).

2. Applications

There are many consequences of 1.3 which we can derive just by considering a very simple (almost degenerate) case of the ordinal games discussed before. Indeed, let λ be an ordinal and given any $R \subseteq \lambda \times \omega^\omega$ associate with it the game:

I ξ — — I plays $\xi < \lambda$ and II follows by playing
 II $\alpha(0)$ $\alpha(1)$... α successively $\alpha(0), \alpha(1), \dots$ from ω ; II wins
 iff $R(\xi, \alpha)$ holds.

If $\varphi: \omega^\omega \rightarrow \lambda$ is a norm, let

$$R^*(w, \alpha) \Leftrightarrow R(\varphi(w), \alpha)$$

be the *coded version* of R . Clearly, the relations R^* as above coincide with the φ -*invariant* relations, where $P(w, \alpha)$ is φ -invariant (on w) if

$$P(w, \alpha) \wedge \varphi(w) = \varphi(v) \Rightarrow P(v, \alpha).$$

By applying 1.3 to the preceding game, we have easily the following uniformization theorem for invariant pointsets.

2.1. THEOREM. *Let $\varphi: \omega^\omega \rightarrow \lambda$ be a Δ_n^1 -norm and $P(w, \alpha)$ a φ -invariant (on w) relation. If $\forall w \exists \alpha P(w, \alpha)$, then there is a Δ_n^1 -function $G: \omega^\omega \rightarrow \omega^\omega$ which uniformizes P , i.e. $\forall w P(w, G(w))$, granting Determinacy ($\text{Proj}[P]$).*

This should be compared with Moschovakis' Main Lemma in MOSCHOVAKIS (1970), where he shows (in this particular case) that in the above situation, there is a *multivalued* Σ_n^1 uniformization for P , i.e. a relation $\bar{P}(w, \alpha)$, which is also φ -invariant, such that $\bar{P} \in \Sigma_n^1$, $\bar{P} \subseteq P$ and $\forall w \exists \alpha \bar{P}(w, \alpha)$.

Actually, a more detailed analysis shows that when $n \geq 2$, given any φ as above there is a Spector pointclass (see MOSCHOVAKIS, 1980) $\Gamma \subseteq \Delta_n^1$ such that G can be taken to be Δ -measurable (i.e. $G^{-1}(N) \in \Delta = \Gamma \cap \neg \Gamma$ for all open sets N). From this and using some standard coding devices for the Δ set of reals by reals, which forms part of the general theory of Spector pointclasses, it is not difficult to give a proof of the following ordinal quantification theorem.

2.2. THEOREM. *Let $\varphi: \omega^\omega \rightarrow \lambda$ be a Δ_n^1 -norm and let $R(w, \alpha)$ be a Σ_{n+1}^1 φ -invariant relation. Then*

$$\exists w R(w, \alpha), \quad \forall w R(w, \alpha)$$

are also Σ_{n+1}^1 , granting PD.

In other words, Σ_{n+1}^1 predicates are closed under quantification over Δ_n^1 ordinals (i.e. lengths of Δ_n^1 prewellorderings of the reals). These ordinals are of course in general uncountable.

Theorem 2.2 can in turn be used to compute a sharp estimate for the complexity of the rank comparison relation between Δ_n^1 prewellorderings of reals (and also wellfounded relations). More specifically, we have:

2.3. THEOREM. *Let $\varphi: \omega^\omega \rightarrow \lambda$, $\psi: \omega^\omega \rightarrow \mu$ be Δ_n^1 -norms. Then the relation*

$$R_{\varphi, \psi}(w, v) \Leftrightarrow \varphi(w) \leq \psi(v)$$

is Δ_{n+1}^1 , granting PD.

We can put now together the computational estimates given in the last two results with Moschovakis' Main Lemma in MOSCHOVAKIS (1970) (or the stronger 2.1) to show that

2.4. THEOREM. *Assume PD, when $n \geq 1$. Then $\omega^\omega \cap L[T^{2n+1}] = C_{2n+2}$.*

Here, of course, T^{2n+1} is the tree coming from a Π^1_{2n+1} -scale on a complete Π^1_{2n+1} set P (see, for example, KECHRIS and MOSCHOVAKIS, 1976–77) and C_{2n+2} the largest countable Σ^1_{2n+2} set. This result has been conjectured by Moschovakis, who proved in KECHRIS and MOSCHOVAKIS (1976–77) the case $n = 0$. The case $n = 1$ was first proved in KECHRIS and MARTIN (1978) (from stronger determinacy hypotheses).

There are further applications of the full length (instead of the ‘one-step’) ordinal games, which will be discussed in the paper mentioned in the introduction. (Also HARRINGTON and SAMI (1979) make use of 1.1 in their work on projective and hyperprojective equivalence relations.) It seems, however, safe to say that the potential of the full ordinal games has not been yet adequately exploited.

Added in proof. Recent results of Martin, Moschovakis and Steel extend the range of applicability of the results and methods in this paper to ordinals much higher than ω^ω . The forthcoming paper of the authors mentioned in the introduction will appear in the Annals of Mathematical Logic under the title *On the determinacy of games on ordinals*.

References

- HARRINGTON, L. A., 1976, *AD(ω) implies a version of AD(\aleph_ω)*, mimeographed note (December)
- HARRINGTON, L. A., and A. SAMI, *Equivalence relations, projective and beyond*, Logic Colloquium 78, M. Boffa, D. Van Dulen, K. McAlloon (eds) (North Holland), pp. 247–264
- KECHRIS, A. S., and D. A. MARTIN, 1978, *On the theory of Π^1_3 sets of reals*, Bulletin of the American Mathematical Society, vol. 84, pp. 149–151
- KECHRIS, A. S., and Y. N. MOSCHOVAKIS, 1976–1977, *Notes on the theory of scales*, Cabal Seminar 76–77, Proceedings of the Caltech-UCLA Logic Seminar, Lecture Notes in Mathematics, vol. 689 (Springer-Verlag)
- MOSCHOVAKIS, Y. N., 1970, *Determinacy and prewellorderings of the continuum*, ed. Y. Bar-Hillel, Mathematical logic and foundations of set theory (North Holland, Amsterdam), pp. 24–62
- MOSCHOVAKIS, Y. N., 1980, *Descriptive set theory* (North Holland)

THE SIMPLEST COUNTEREXAMPLE TO COMPACTNESS IN THE CONSTRUCTIBLE UNIVERSE

MENACHEM MAGIDOR

Institute of Mathematics, Hebrew University, Jerusalem, Israel

0. Introduction

The compactness referred in the title of this paper is a generalization of the Barwise compactness theorem (BARWISE, 1969), i.e. given an admissible set A , and a set B where $B \subseteq L_{\omega\omega} \cap A$, we expect B to have a model if every A finite subset of B has a model. ("A finite" means being a member of A .) Admissible sets for which this compactness property holds for every B are rather difficult to come by. For instance, assuming $V = L$, and considering only admissible sets of the form L_α , the only α 's such that we have the full compactness are the weakly compact cardinals. Hence for most of α 's we can find a set $B \subseteq L_{\omega\omega} \cap L$ such that every A finite subset of B has a model but B does not have a model. The theme of this paper is finding the "simplest" such counterexamples to compactness.

The answer to our problem depends of course on the notion "simple". Throughout the paper we assume the axiom of constructibility ($V = L$). We shall say that a set C is *simpler* than B if C is constructed before B in the usual procedure for generating the constructible universe, i.e. if an ordinal γ such that $C \in L_{\gamma+1} - L_\gamma$ is smaller than an ordinal δ such that $B \in L_{\delta+1} - L_\delta$. We are not distinguishing between sets constructed at the same level, though a finer version of our results are possible if one considers a finer structure of the constructible hierarchy in the sense of JENSEN (1972). (See also DEVLIN, 1973.)

Let $\gamma(\alpha)$ be the first ordinal γ such that in $L_{\gamma+1}$ we have a counterexample to the compactness of L_α . Let $\beta(\alpha)$ be the first ordinal β (if it exists) such that in $L_{\beta+1}$ one can find a set $B \subseteq \alpha$ witnessing the fact that α is singular, i.e. B is a subset of α of order type less than α , cofinal in α . $\delta(\alpha)$ is the first

ordinal δ such that in $L_{\delta+1}$ one can find a stationary subset of α with no stationary initial segment. What we are interested in is a characterizing of $\gamma(\alpha)$. For regular α 's, the problem is essentially solved in JENSEN (1972), where it is shown that if α is regular, then $\delta(\alpha) = \gamma(\alpha)$. On the other hand, it follows from the Barwise compactness theorem that for countable α 's $\beta(\alpha) = \gamma(\alpha)$. (It applies to other α 's as well provided they satisfy the appropriate generalization of the Barwise theorem. See MAGIDOR, SHELAH, and STAVI, forthcoming.) Note that for α of cofinality ω $\delta(\alpha)$ does not make much sense.

So our attention is drawn to the singular α 's which are uncountable. In $\beta(\alpha)+1$ we have an obvious counterexample to the compactness of α . Hence $\gamma(\alpha) \leq \beta(\alpha)$ is true for every α . Is $\gamma(\alpha) = \beta(\alpha)$ for every singular α ? The following theorem gives a negative answer.

THEOREM 1. ($V = L$) Let $\kappa > \omega_1$ be a regular cardinal; the set

$$\{\alpha \mid \gamma(\alpha) = \beta(\alpha)\}$$

is both stationary and co-stationary in κ .

Theorem 1 is proved in Section 1.

We are looking for a combinatorial characterization of $\gamma(\alpha)$ and in Section 2 we give a partial answer for a large class of α 's. In order to state our results we need some notation. $P_\kappa(\lambda)$ is the set of all subsets of λ of cardinality less than κ . $A \subseteq P_\kappa(\lambda)$ is closed if A is closed under unions of increasing chains of length less than κ . $A \subseteq P_\kappa(\lambda)$ is unbounded if for every $P \in P_\kappa(\lambda)$ $\exists Q \in A$ such that $Q \supseteq P$. $A \subseteq P_\kappa(\lambda)$ is stationary in $P_\kappa(\lambda)$ if it intersects every closed unbounded subset of $P_\kappa(\lambda)$. (See JECH, 1973.) Let $\eta(\alpha)$ be the first ordinal η such that in $L_{\eta+1} - L_\eta$ one can find a stationary subset of $P_{\omega_1}(\alpha)$, A , such that for every $\beta < \alpha$ $A \cap P_{\omega_1}(\beta)$ is non-stationary in $P_{\omega_1}(\beta)$.

THEOREM 2. ($V = L$) Let $|\alpha| = \kappa$ where $\text{cf}(\kappa) > \omega$, α admissible, $\text{cf}(\alpha) > \omega$. Assume also that $L_\alpha \models \kappa$ is the last cardinal. Then $\gamma(\alpha) = \eta(\alpha)$.

A simpler version of Theorem 2 is obtained from Theorem 2 by using methods of LITMAN.

THEOREM 3. ($V = L$) Let $\text{cf}(\alpha) = \kappa$, $|\alpha| = \kappa$, α admissible with the additional assumption that $\kappa = \delta^+$ for δ regular. $L_\alpha \models \kappa$ is the last cardinal; then $\gamma(\alpha) = \delta(\alpha)$.

The proof of Theorem 3 will not be given here. We refer the reader to a future publication of Litman.

1. $\beta(\alpha)$ is not always the right answer

In this section we prove Theorem 1. We have two clauses to verify:

- (A) $\{\alpha \mid \alpha < \kappa, \beta(\alpha) = \gamma(\alpha)\}$ is stationary in κ .
- (B) $\{\alpha \mid \alpha < \kappa, \beta(\alpha) > \gamma(\alpha)\}$ is stationary in κ .

(A) is an easy consequence of the following lemma due to Jensen.

LEMMA 4. ($V = L$.) Let κ be regular. The set B of all $\alpha < \kappa$ such that $\beta(\alpha) = \gamma + 1$ for some γ such that there exists an elementary embedding $j: \langle L_\gamma, \in \rangle$ into $\langle L_{\kappa^+}, \in \rangle$, $j(\delta) = \delta$ for $\delta < \alpha$, $j(\alpha) = \kappa$, is stationary in κ .

Let $\alpha \in B$. We claim that $\gamma(\alpha) = \beta(\alpha)$. Assume otherwise. Hence in $L_{\beta(\alpha)}$ there exists a counterexample to the compactness of L_α , A . (Note: $\gamma(\alpha) \leq \beta(\alpha)$ is always true.) A is a subset of $L_{\omega\omega} \cap L_\alpha$ with no model, but for $\gamma < \alpha$ $A \cap L_\gamma$ has a model.

Let $j: L_\delta \rightarrow L_{\kappa^+}$ where $\delta + 1 = \beta(\alpha)$ and j is an elementary embedding fixing every member of α (hence, every member of L_α) and $j(\alpha) = \kappa$. Note that A is first order definable over L_δ with a parameter $p \in L_\delta$. Apply the same definition in L_{κ^+} to $j(p)$ to get a set D . Since j is an elementary embedding, D is a subset of $L_{\omega\omega} \cap L_{j(\alpha)}$ ($j(\alpha) = \kappa$). Since j is the identity on L_α , we also get $D \cap L_\alpha = A$. But A does not have a model. Hence A does not have a model in L_{κ^+} . Therefore,

$$L_{\kappa^+} \models \exists \gamma < \kappa \ (D \cap L_\gamma \text{ has no model}).$$

Therefore, by j being elementary,

$$L_\delta \models \exists \gamma < \alpha \ (A \cap L_\gamma \text{ has no model}).$$

Fix γ satisfying the last formula. Note that $j(A \cap L_\gamma) = A \cap L_\gamma$. (Since $\gamma < \alpha$.) Hence

$$L_{\kappa^+} \models A \cap L_\gamma \text{ has no model}.$$

The last conclusion clearly implies that $A \cap L_\gamma$ has no model since any such model should lie in L_{κ^+} . This is a contradiction to our assumption that $A \cap L_\gamma$ has a model for every $\gamma < \alpha$ and (A) is verified.

Now we prove (B). Namely, we prove that the set

$$C = \{\alpha \mid \alpha < \kappa, \gamma(\alpha) < \beta(\alpha)\}$$

is also stationary in κ . Note that we make the implicit assumption that $\alpha \in C$ implies that $\beta(\alpha)$ exists, i.e. that α is singular.

The proof is divided into two parts. In the first part we prove the theorem for κ which is not Mahlo (κ is Mahlo = The set of regulars is stationary in κ), and then we shall deduce that the theorem also holds for κ which is Mahlo. So assume until further notice that κ is not Mahlo, i.e. there exists a closed unbounded subset of κ of singular ordinals.

We shall define a set $A \subseteq \kappa$ by induction. Given $A \cap \alpha$, we have to decide whether $\alpha \in A$ or not and we distinguish two cases.

(i) $\omega < \text{cf}(\alpha) < \alpha$ and for all $\gamma < \alpha$, $A \cap \gamma \in L_\alpha$, $A \cap \alpha \in L_{\beta(\alpha)}$,

$$L_\alpha \models \forall \beta \in A \cap \alpha \ \beta \text{ is singular},$$

$L_{\beta(\alpha)} \models \{\beta \mid \beta < \alpha \text{ is regular}\}$ is non-stationary in α ,

$$L_{\beta(\alpha)} \models A \cap \alpha \text{ is non-stationary}.$$

In this case we make $\alpha \in A$.

(ii) Case (i) fails. In this case we make $\alpha \notin A$.

LEMMA 5. *A is a stationary subset of κ .*

PROOF: Assume otherwise. In L_{κ^+} one can find a closed unbounded subset of κ , D , disjoint from A . Since κ is not Mahlo, we can assume without loss of generality that every member of D is singular. Let N be an elementary substructure of $\langle L_{\kappa^+}, \varepsilon, A, D \rangle$ of cardinality $< \kappa$ which satisfies $\alpha = N \cap \kappa \in \kappa$ for some $\alpha \in \kappa$ of cofinality $> \omega$.

N is isomorphic to a structure of the form $\langle L_\beta, \varepsilon, A \cap \alpha, D \cap \alpha \rangle$ where the image of κ is α . Note that α is regular in L_β ; hence $\beta < \beta(\alpha)$. Therefore, $D \cap \alpha \in L_{\beta(\alpha)}$, $A \cap \alpha \in L_{\beta(\alpha)}$ and since α is the image of κ for $\gamma < \alpha$, we have $A \cap \gamma \in L_\alpha$, and for $\alpha \in A \cap \alpha$ $L_\alpha \models \gamma$ is singular. $D \cap \alpha$ is unbounded in α ; hence $\alpha \in D$. Therefore α is singular. We proved that case (i) of the definition of A applies to α , therefore $\alpha \in A$ but this is a contradiction to $\alpha \in D$. ■ (Lemma 5).

We shall finish the proof of Theorem 2 (at least for the case κ not Mahlo) if we show that $A \subseteq C$.

Let $\alpha \in A$. By the definition of A we can find in $L_{\beta(\alpha)}$ a closed unbounded subset of α , D , such that every member of D is singular in L_α and D is disjoint from A . Also: $A \cap \alpha \in L_{\beta(\alpha)}$. We shall show how one can define from D and $A \cap \alpha$ a theory $T \subseteq L_{\infty\omega} \cap L_\alpha$ which is a counterexample to the compactness of L_α . The definition is simple enough so that it will be clear that since D and $A \cap \alpha$ are in $L_{\beta(\alpha)}$, T is in $L_{\beta(\alpha)}$.

Let T_α be a theory whose language is the language of set theory with the addition of a constant c_x for every $x \in L_\alpha$ as well as an additional constant d .

The axioms of T_α are made up of the following kinds:

- (i) $KP + V = L$.
- (ii) $\forall x(x \in c_y \leftrightarrow \bigvee_{z \in y} z = c_z)$ for all $y \in L_\alpha$.
- (iii) d is an ordinal.
- (iv) $c_\gamma < d$ for all $\gamma < \alpha$.

T_α is a subset of $L_{\omega\omega} \cap L_\alpha$ since α is admissible. Note that T_α is Σ_1 over L_α . T_α is a part of our final theory T , T_α guarantees that a model of T is a proper end extension of $\langle L_\alpha, \in \rangle$.

T contains two additional unary predicates \tilde{D} , \tilde{A} with the additional axioms:

- (v) ' \tilde{D} is a closed unbounded class of singular ordinals.'
- (vi) ' \tilde{A} is a class of ordinals, α , for which case i of the definition of A holds with $\tilde{A} \cap \alpha$ replacing $A \cap \alpha$ '
- (vii) $\tilde{D} \cap \tilde{A} = \emptyset$.
- (viii) $\tilde{A}(c_\beta)$ for $\beta \in A$ and $\neg \tilde{A}(c_\beta)$ for $\beta \notin A$.
- (ix) $\tilde{D}(c_\beta)$ for $\beta \in D$ and $\neg \tilde{D}(c_\beta)$ for $\beta \notin D$.

If $\beta < \alpha$ then clearly $T \cap L_\beta$ has a model $\langle L_\alpha, \in, A, D \rangle$ is such a model where c_x for $x \in L_\alpha$ is interpreted as x and d as any ordinal in L_α greater than any ordinal mentioned in $T \cap L_\beta$. Theorem 2 for the non-Mahlo cardinal will be established if we prove:

LEMMA 6. T has no models.

PROOF: Assume otherwise and let $\langle M, E, \tilde{A}, \tilde{D} \rangle$ be such a model. M has an initial segment which is well founded; hence we can assume that it has the form $\langle L_\gamma, \in \rangle$. The fact that M is a model of T implies that $\alpha \leq \gamma$. The first claim is that without loss of generality we can have: $\alpha < \gamma$. Assume otherwise. Denote the realization of the constant d also by d . Since $\gamma = \alpha$, d is already in the non-standard part of M . Let N be the Σ_1 Skolem closure in M (note that since $M \models V = L$, we have definable Σ_1 Skolem function) of $L_\alpha \cup \{d\}$. (Using also the predicates \tilde{A} and \tilde{D} .)

The following lemma is essentially proved in MAGIDOR, SHELAH, and STAVI (forthcoming); since we are not sure when the paper will appear, we shall supply the proof of Lemma 7.

LEMMA 7. α is in the well founded proof of N .

Assume Lemma 7. The proof will be supplied later. N can easily be expanded to a model of T (Note: $L_\alpha \cup \{d\} \subseteq N$) and in view of Lemma 7 the well founded part of N includes α ; hence we can assume $\alpha < \gamma$.

Since $\tilde{D} \cap \alpha = D \cap \alpha$, \tilde{D} is unbounded in α . Hence, in view of axiom v of T , $\alpha \in \tilde{D}$ and α is singular in N . Another lemma from MAGIDOR, SHELAH, and STAVI (forthcoming) which we shall use is the following

LEMMA 8. *Let $\langle M, E \rangle$ be a model of $KP + V = L$. Let α be an ordinal in the standard part of $\langle M, E \rangle$ with $\text{cf}(\alpha) > \omega$. If $\langle M, E \rangle \models \alpha$ is singular, then $\beta(\alpha)$ is in the standard part of $\langle M, E \rangle$.*

SKETCH OF THE PROOF: It follows from JENSEN (1972) (see also DEVLIN 1973) that $L_{\beta(\alpha)}$ can be represented as a direct limit of structures of the form $\langle L_\gamma, \in \rangle$ for $\gamma < \alpha$ where the cofinality of the directed system is $\text{cf}(\alpha)$. Apply this fact in $\langle M, E \rangle$ and get $\beta^M(\alpha)$ ($\beta(\alpha)$ in the sense of M) as a direct limit of well founded structures. Since the cofinality of this directed system is really $> \omega$ (since $\text{cf}(\alpha) > \omega$), $\beta^M(\alpha)$ is well founded and it is in the standard part of $\langle M, E \rangle$. ■ (Lemma 8)

In view of Lemma 8, $\beta(\alpha) < \gamma$. Remember that $\alpha \in A$. It is easy to check that since $A \cap \alpha = \tilde{A} \cap \alpha$ and $\beta(\alpha) < \gamma$, we have $\alpha \in \tilde{A}$. (Remember axiom vi.) Hence $\alpha \in \tilde{A} \cap \tilde{D}$ which is a contradiction and we have proved. ■ (Lemma 6 (modulo Lemma 7)).

PROOF OF LEMMA 7: We prove a claim from which the lemma will follow.

CLAIM. *For every $\delta < \alpha$ and every finite $p \subseteq M$, let $R(p, \delta)$ be the Σ_1 closure in M of $L_\delta \cup p$; then*

- (a) *$R(p, \delta)$ contains an M ordinal which is E minimal in $R(p, \delta) - L_\alpha$ (provided this last set is not empty).*
- (b) *$R(p, \delta)$ is bounded in L_α .*

PROOF OF THE CLAIM: The first thing to observe is that for all δ (b) implies (a), because for $\delta < \alpha$ one can define $R(\delta, p)$ in M ; hence if $R(\delta, p) \cap \alpha < \gamma$ for some γ , there exists the minimal ordinal in $R(\delta, p) - L_\gamma$. Hence (a) is verified provided (b) is given.

We prove (a) and (b) by induction on $\delta < \alpha$. For $\delta = 0$, (b) is trivial since $R(0, p)$ is countable and $\text{cf}(\alpha) > \omega$. Hence $R(0, p) \cap \alpha$ is bounded in α . Hence (a) is verified.

For $\delta + 1$, (a) and (b) easily follow from the claim for δ since $R(\delta + 1, p) = R(\delta, p \cup \{\delta\})$. For δ limit of cofinality ω , (b) is clear since $R(\delta, p) = \bigcup_{\mu < \delta} R(\mu, p)$, and since by induction assumption $R(\delta, p)$ is bounded in α , $R(\delta, p)$ is bounded in α , because $\text{cf}(\delta) = \omega$ but $\text{cf}(\alpha) > \omega$.

For δ limit of cofinality $> \omega$, we verify (a) first. By induction assumption we have for every $\mu < \delta$ a minimal ordinal in $R(\mu, p) - L_\alpha$. Denote this

minimal ordinal by d_μ . Clearly, for $\mu < \mu^1$, $d_{\mu^1} Ed_\mu$ or $d_{\mu^1} = d_\mu$. We claim that $\{d_\mu \mid \mu < \delta\}$ are well founded, otherwise if $\{d_\mu \mid i < \omega\}$ is a descending sequence, we have $\mu_i < \mu_{i+1}$. Let $\mu = \sup_{i < \omega} \mu_i$. Since $\text{cf}(\delta) > \omega$, we have $\mu < \delta$. But $R(\mu, p) = \bigcup_{i < \omega} R(\mu_i, p)$. Hence $d_\mu \in R(\mu_i, p)$ for some i . This clearly contradicts $d_{\mu_{i+1}} Ed_{\mu_i}$. We got a contradiction and $\{d_\mu \mid \mu < \delta\}$ is well founded. Let d be its minimal element. Since $R(\delta, p) = \bigcup_{\mu < \delta} R(\mu, p)$, d is the minimal element in $R(\delta, p) - L_\alpha$ and (a) of the Claim is verified for this last case. (b) follows from (a), because if $R(\delta, p)$ is unbounded in L_δ , in M one can find an ordinal l which is the supremum of all ordinals in $R(\delta, p)$ less than d , where d is the minimal ordinal in $R(\delta, p) - \alpha$ which exists by (a). l is clearly the bound of $\{\beta \mid \beta < \alpha\}$ and clearly $\{f \mid f El\}$ has order type α ; hence α is in the well founded part of M which contradicts our assumptions. ■ (Claim)

Lemma 7 follows from the Claim by noting that N is $R(\alpha, p) = \bigcup_{\delta < \alpha} R(\delta, p)$.

The reason that $R(\alpha, p) - \alpha$ has a minimal ordinal is exactly like proving (a) of the Claim in the case $\text{cf}(\delta) > \omega$, using the claim for $\delta < \alpha$. ■ (Lemma 7)

Thus we have verified Theorem 1 for κ which is not Mahlo. For κ which is Mahlo we use:

LEMMA 9. *If κ is Mahlo, then the set $A = \{\alpha \mid \alpha \text{ is regular but not Mahlo}\}$ is stationary in κ .*

PROOF: Assume otherwise; hence there exists a closed unbounded subset of κ , D , disjoint from A . Since κ is Mahlo, D contains some regular limit points. Let α be the first regular limit point of D . $D \cap \alpha$ is a closed unbounded subset of α with no regular limit points. Hence α is not Mahlo, hence $\alpha \in A \cap D$. A contradiction. ■ (Lemma 9)

Note that by our previous proof $C \cap \alpha$ is stationary in α for every $\alpha \in A$, but this clearly implies that C is stationary in κ . Otherwise, if D is closed unbounded and disjoint from C , then by Lemma 9 D has a limit point in A , α . $D \cap \alpha$ proves first $C \cap \alpha$ is not stationary in α . A contradiction. We verified Theorem 1 for the case where κ is Mahlo. ■ (Theorem 1)

2. So what is $\gamma(\alpha)$?

In this section we prove Theorem 2. The crucial tool is the Kueker approximation of a theory in $L_{\omega\omega}$. Our definition is a minor modification of Kueker's one, though equivalent to it for all practical matters.

Let $T \subseteq L_{\omega\omega} \cap L_\alpha$. Let $P \in P_{\omega_1}(L_\alpha)$. For a closed unbounded set of P 's in $P_{\omega_1}(L_\alpha)$, $\langle P, \epsilon \rangle$ is an extensional structure; hence it is isomorphic to a countable transitive set. Let h_P be this isomorphism. It is easy to see that for a closed unbounded subset of P 's in $P_{\omega_1}(L_\alpha)$, the image of h_P is exactly $L_{|P|}$ where $|P|$ is the order type of $P \cap \alpha$. For P in this closed unbounded set, T_P , the P th Kueker approximation, is $h_P''(T \cap P)$ and again for a closed unbounded set of P 's $T_P \subseteq L_{\omega\omega} \cap L_{|P|}$.

KUEKER (1977) proved that if T has a model, then for a closed unbounded set of P 's in $P_{\omega_1}(L_\alpha)$ T_P has a model. The following lemma includes a kind of converse to Kueker's Theorem in $V = L$.

Notation: For $P \in P_{\omega_1}(L_\alpha)$, let $\varrho(P)$ be the first ordinal in which $|P|$ is countable.

LEMMA 10. ($V = L, \alpha \geq \omega_1$.) $T \subseteq L_{\omega\omega} \cap L_\alpha$ has a model if and only if the set $\{P \mid P \in P_{\omega_1}(L_\alpha), T_P \text{ has a model in } I_{\varrho(P)}\}$ contains a closed unbounded set.

PROOF: *Only if:* Consider the structure $M = \langle L_{\alpha^+}, \epsilon, T \rangle$. If T has a model, it has a model in L_{α^+} . The set of all Q 's in $P_{\omega_1}(L_{\alpha^+})$ such that $\langle Q, \epsilon, Q \cap T \rangle$ is an elementary substructure of M , and contains a closed unbounded subset of $P_{\omega_1}(L_{\alpha^+})$. Denote such a closed unbounded set by B . For $Q \in B$ $\langle Q, \epsilon, T \cap Q \rangle$ is isomorphic to a structure of the form $\langle L_\beta, \epsilon, T_{Q \cap L_\alpha} \rangle$, where $T_{Q \cap L_\alpha}$ is the appropriate Kueker approximation of T . Since $L_{\alpha^+} \models T$ has a model, we get $L_\beta \models T_{Q \cap L_\alpha}$ has a model. Note that since α is uncountable in L_β , $|Q \cap \alpha|$ is uncountable in L_β . Hence $\beta < \varrho(Q \cap L_\alpha)$, and $T_{Q \cap L_\alpha}$ has a model in $L_{\varrho(Q \cap L_\alpha)}$. The set $C = \{P \mid P = Q \cap L_\alpha \text{ for some } Q \in B\}$ clearly contains a closed unbounded subset of L_α . Hence the *only if* direction of our lemma is verified.

If direction: Assume that T does not have a model. Hence it does not have a model in $\langle L_{\alpha^+}, \epsilon \rangle$. By an argument completely analogous to the proof of Lemma 4 in JENSEN (1972), one can prove that the set $C = \{P \in P_{\omega_1}(L_{\alpha^+}) \mid \varrho(P) \text{ is a successor ordinal of the form } \delta(P)+1 \text{ such that } L_{\delta(P)} \text{ can elementarily be embedded in } L_{\alpha^+} \text{ by an embedding } j_P \text{ satisfying } j_P \upharpoonright L_{|P|} = h_P^{-1}, \text{ and } j(T_P) = T\}$ is a stationary subset of $P_{\omega_1}(L_\alpha)$. For P in C , T_P does not have a model in $L_{\varrho(P)}$, because such a model would be a definable subset of $L_{\delta(P)}$, and j_P applied to this model would yield a model of T in L_{α^+} . We proved that the set of P 's such that T_P does not have a model in $L_{\varrho(P)}$ is stationary in $P_{\omega_1}(L_\alpha)$. Therefore the complement of this set cannot contain a closed unbounded subset. ■ (Lemma 10)

PROOF OF THEOREM 2: Let $|\alpha| = \kappa$ where $\text{cf}(\alpha) > \omega$ and such that $L_\alpha \models \kappa$ is the last cardinal. Note that our assumption guarantees that, for every $P \in P_{\omega_1}(\alpha)$, $P \in L_\alpha$. Note also that in every $L_\alpha (\alpha \geq \omega)$ one can define a one-to-one function from α onto L_α . Denote such a function by f . We prove first $\eta(\alpha) \leq \gamma(\alpha)$. Let T be a counterexample to the compactness of L_α . Hence T does not have a model, whereas for every $\beta < \alpha$ $T \cap L_\beta$ does have a model. Let $r(x)$ be a minimal β such that $x \in L_\beta$. Consider the set:

$$A = \{P \mid P \in P_{\omega_1}(\alpha), P \text{ is closed under } r \circ f, T_{f''P} \text{ does not have a model in } L_{\eta(f''P)}\}.$$

In view of Lemma 10, A is a stationary subset of $P_{\omega_1}(\alpha)$. For $\beta < \alpha$ $A \cap P_{\omega_1}(\beta)$ is non-stationary, because $T \cap L_\beta$ has a model and if $P \in A \cap P_{\omega_1}(\beta)$ then $f''P \subseteq L_\beta$. (Lemma 10 is used again.) Since A is constructed at the same stage at which T is constructed, we have proved $\eta(\alpha) \leq \gamma(\alpha)$.

Now we prove $\gamma(\alpha) \leq \eta(\alpha)$. Suppose that at a certain stage we constructed a stationary subset of $P_{\omega_1}(\alpha)$ such that for all $\beta < \alpha$ $A \cap P_{\omega_1}(\beta)$ is non-stationary; then we show that at the same stage we constructed theory $T \subseteq L_{\omega_\omega} \cap L_\alpha$ which is a counterexample to the compactness of L_α .

The theory T will be the union of the theory T_α defined in the proof of Lemma 5 and a theory containing the additional symbols: F , which is a unary function symbol, a Q , a unary predicate, and G, H which are binary functions and the additional axioms:

- (1) F is a one-to-one function from the domain of the model onto c_κ .
- (2) $\forall \alpha < c_\kappa G(\alpha, \cdot)$ is an isomorphism of ordinals in $F^{-1}(\alpha)$ onto δ for some $\delta < \kappa$.

(1)+(2) are essentially the usual axiom one has when one applies Chang's trick for guaranteeing that the resulting model will be well founded using $\text{cf}(\kappa) > \omega$.

- (3) $KP + V = L$.
- (4) $\forall x Q(x) \rightarrow x$ is countable.
- (5) $\forall \beta H(\beta, \cdot)$ is a function defined on a finite subset of β such that every countable subset of β closed under $H(\beta, \cdot)$ is not in Q .
- (6) For every $P \in A$ we have an axiom ' $Q(C_P)$ ' and for every $P \in L_\alpha - A$ we have ' $\neg Q(C_P)$ '.

For every $\beta < \alpha$, $T \cap L_\beta$ has a model. $\langle L_\alpha, \varepsilon, A \rangle$ can be expanded to such a model by interpreting d as any ordinal greater than any ordinal mentioned in $T \cap L_\beta$. F and G can easily be found by $|\alpha| = \kappa$. Appropriate H can be

found by the fact that $A \cap P_{\omega_1}(\beta)$ is non-stationary for $\beta < \alpha$, and hence its complement contains a closed unbounded subset of $P_{\omega_1}(\beta)$. It is well known that any such closed unbounded set contains a set of the form

$$\{P \mid P \in P_{\omega_1}(\beta), P \text{ is closed under } F\},$$

where F is some function from $P_\omega(\beta)$ into β , $H(\beta, \cdot)$ is picked to be any such function.

T does not have a model, because such a model would have to be well founded and an end extension of L_α . Hence it should contain the ordinal α . $H(\alpha, \cdot)$ verifies that $Q \cap P_{\omega_1}(\alpha)$ is not stationary, but $Q \cap P_{\omega_1}(\alpha)$ is A , and we get a contradiction. ■ (Theorem 2).

References

- BARWISE, J., 1969, *Infinitary logic and admissible sets*, Journal of Symbolic Logic, vol. 36, pp. 226–252
- DEVLIN, K., 1973, *Aspects of constructibility*, Lecture notes in mathematics, 354 (Springer-Verlag, Berlin, Heidelberg, New York)
- JECH, T. J., 1973, *Some combinatorial problems concerning uncountable cardinals*, Annals of Math. Logic, vol. 5, pp. 165–198
- JENSEN, R. B., 1972, *The fine structure of the constructible hierarchy*, Annals of Math. Logic, vol. 4, pp. 239–308
- KUEKER, D. W., 1977, *Countable approximation and Lowenheim-Skolem theorems*, Annals of Math. Logic, vol. 11, pp. 57–103
- LITMAN, A., *Properties of L*, Ph. D. Thesis, Hebrew University, Jerusalem
- MAGIDOR, M., S. SHELAH, J. STAVI, *Admissible set theory at cofinality ω* (in preparation)

ANDERSON AND BELNAP, AND LEWY ON ENTAILMENT*

J. MICHAEL DUNN

Indiana University. Bloomington, Indiana, U.S.A.

ANDERSON and BELNAP (1962, p. 47) seem to feel that it is a mistake "to try to build a sieve which will 'strain out' entailments from the set of material or strict 'implications' present in some system of truth-functions, or of truth-functions with modality." In this paper I show that just such a "sieve" can be constructed, adapting an early idea of LEWY (1958, more explicitly in 1976). This sieve in fact gives just the first-degree entailments (entailments between truth-functions as opposed to entailments between entailments) of the Anderson-Belnap system **E** of entailment.

In the way of background, it should be remarked that in the late 1950's an interesting proposal concerning entailment developed in the work of von Wright, Geach, and Smiley. This proposal is formulated in ANDERSON and BELNAP (1975) as the

WGS CRITERION. *A entails B iff (i) $A \supset B$ is a substitution instance of a tautology $A' \supset B'$, where (ii) A' is not a contradiction and (iii) B' is not a tautology.*

* This is a resume of the lecture given under the title *A sieve for entailments*. A more rigorous and more scholarly cousin is to be found under that same title in *Journal of Philosophical Logic*, Feb. 1980. The present resume stresses key ideas by developing a suggestion (encapsulated in Theorem 1 below) not found in the JPL paper, and then saying how it is straightforward to produce a new proof of the principal result of the JPL paper (Theorem 2 below) by making connections to this suggestion. The reader should also consult the JPL paper for a more leisurely connection to the original ideas of Lewy, for connections to the work of others (esp. M. Clark and A. Urquhart), and for thanks (to G. Hunter, C. Lewy, and T. J. Smiley).

This criterion deals handily with such notorious paradoxes of entailment as

- (1) $p \rightarrow (q \vee \sim q)$,
- (2) $(p \& \sim p) \rightarrow q$,

but it is too timid with respect to such close cousins as

- (3) $p \rightarrow [p \& (q \vee \sim q)]$,
- (4) $[(p \& \sim p) \vee q] \rightarrow q$.

LEWY (1958, 1976) is very much bothered by (3), which he thinks is ‘very nearly, if not quite’ as counterintuitive as (1), and indeed (3), together with the obviously acceptable

- (5) $[p \& (q \vee \sim q)] \rightarrow (q \vee \sim q)$,

would lead by transitivity to (1), which illustrates the well-known lack of transitivity of the WGS entailment. There is of course a dual point about (4) and (2).

We might take Lewy’s point to be that although in checking the entailment (3) by verifying the tautologyhood of

- (3') $p \supset [p \& (q \vee \sim q)]$

we do not thereby discover the tautologyhood of the consequent (as we would in so checking (1)), we still come to discover that the consequent is a “partial tautology.” It has after all a tautologous conjunct, which plays an untoward role in making (3’) a tautology (much like the role of $q \vee \sim q$ with respect to (1)).

In this paper I attempt to frame a ‘correct’ definition of the notion of a “partial tautology,” and the dual notion of a “partial contradiction.” LEWY (1958, 1976) tried to do a similar thing, but he never explicitly invoked the terminology of “partial tautologyhood,” but in its place used the notion of “pure contingency.” In DUNN (1980) I take great pains to argue that this is simply an accident of exposition and that there is an essential conceptual equivalence between Lewy’s literal approach and the approach taken here, modulo some amendments which will become plain in a moment.

Postponing for the moment then the precise definitions, let us state the basic idea of Lewy’s conception as the

L CRITERION. *A entails B iff (i) $A \supset B$ is a substitution instance of a tautology $A' \supset B'$, where (ii) A' is not a “partial contradiction” and (iii) B' is not a “partial contradiction.”*

Turning now to Lewy's own proposal as to how in effect to define the notion of a “partial tautology,” the suggestion is that we look to see whether a given sentence A is “equivalent” in some sense to some conjunction with a tautologous conjunct. (The dual notion of a “partial contradiction” would involve A being “equivalent” to some disjunction with a contradictory disjunct.) The problem is that ordinary logical equivalence will not do because of the tautology

$$(6) A \equiv A \& (q \vee \sim q).$$

Lewy proposed instead that for A to be ‘equivalent’ to B it be required that (i) there exists some tautology $A' \equiv B'$ such that $A \equiv B$ is a substitution instance of it, and (ii) that no proper subformula of $A' \equiv B'$ is tautologous. Requirement (ii) is clearly aimed at (6), but it misfires in virtue of the tautology

$$(7) p \equiv p \& (p \vee q),$$

which has as a substitution instance (i)

$$(8) A \equiv A \& (A \vee \sim A).$$

Now I have two suggestions as to how Lewy's notions should be fixed. Suggestion 1 is not the one which I prefer, but there is some heuristic value in my starting with it. The idea is to define A to be “equivalent” to B iff A can be transformed into B by means of the normal form rules: commutation, association, idempotence, distribution, double negation, and de Morgan's laws (but obviously not such an equivalence as (6) or its dual, which are sometimes used in obtaining “normal forms”).

I think that people who have reflected on the matter would tend to agree that there is a certain naturalness in choosing the normal form rules. Indeed, they have a good claim for generating a natural equivalence relation of synonymy (in at least some sense) at the level of truth-functional logic, and this *a fortiori* gives them a good claim for constituting the “equivalence” relation needed in defining “partial tautology (contradiction).” But still I feel that perhaps there is something *ad hoc* in taking just these

rules (and no more), and this is why I prefer suggestion 2, to which we shall eventually turn.

For now we continue to discuss suggestion 1, relating it to the Anderson-Belnap idea of ‘tautological entailment’ (cf. ANDERSON and BELNAP, 1975, § 15). Where A and B are sentences containing only the connectives $\&$, \vee , and \sim (no \rightarrow), $A \rightarrow B$ is said to be a *tautological entailment* iff, when A has been put in disjunctive normal form $A_1 \vee \dots \vee A_m$ and B has been put in conjunctive normal form $B_1 \& \dots \& B_n$ (using the normal form rules of suggestion 1), then each A_i shares some *atom* ($=_{\text{df}}$ some sentential variable or its negate) with each B_j . Anderson and Belnap have shown that $A \rightarrow B$ is a tautological entailment iff $A \rightarrow B$ is a theorem of their system **E** (A, B with no \rightarrow).

It is now almost immediate to connect the first-degree entailments of **E** with the entailments according to the L Criterion (with this last spelled out according to suggestion 1). Thus let us suppose that $A \rightarrow B$ is a tautological entailment. A typical example might look like this:

$$(9) \sim p \& \sim [\sim p \& \sim (q \& r)] \rightarrow p \vee [r \& (q \vee \sim q)].$$

Putting the antecedent in disjunctive normal form and the consequent in conjunctive normal form then gives

$$(9\text{nf}) (p \& \sim p) \vee (\sim p \& q \& r) \rightarrow (p \vee r) \& (p \vee q \vee \sim q)$$

(which can easily be seen to pass the Anderson–Belnap component-wise sharing test). The problem is, as (9nf) clearly displays, that the antecedent of (9) is a partial contradiction, and the consequent is a partial tautology. The solution is to rewrite certain occurrences of p and q in (9nf) as new sentential variables, say p' and q' , so as to destroy the partial contradictoriness of its antecedent and the partial tautologyhood of its consequent, obtaining thereby

$$(9\text{nf}') (p' \& \sim p) \vee (\sim p \& q' \& r) \rightarrow (p' \vee r) \& (p' \vee q' \vee \sim q)$$

(shared occurrences must also be rewritten so as to preserve tautological entailmenthood).

It is then clear that the normal form steps that produced (9nf), when applied in reverse order to (9nf') will produce

$$(9') \sim p \& \sim [\sim p' \& \sim (q' \& r)] \rightarrow p' \vee [r \& (q' \vee \sim q)],$$

which has (9) as a substitution instance. (9') is a tautology (putting \supset for \rightarrow); indeed, because component-wise sharing has been maintained, it is a tautological entailment. So condition (i) of the L Criterion is met. And it is of course true that conditions (ii) and (iii) are met as well, because of the business of rewriting p and q (in some occurrences).

The procedure we have just illustrated is perfectly general, as the reader can convince himself by working through the example above and perhaps some others, and I shall not provide here the inductions needed to be rigorous.

Attacking now the converse, let us assume that $A \rightarrow B$ passes the L Criterion and show that it is a tautological entailment. Then (i) $A \rightarrow B$ is a substitution instance of a tautology $A' \supset B'$. Putting A' into disjunctive normal form $A'_1 \vee \dots \vee A'_m$, then (ii) no A'_i is a contradiction. Putting B' into conjunctive normal form $B'_1 \& \dots \& B'_n$, then (iii) no B'_j is a tautology. It is clear then (cf. ANDERSON and BELNAP, 1975, § 15) that since $A' \supset B'$ is a tautology, then each $A'_i \supset B'_j$ is a tautology. And since no A'_i is a contradiction and no B'_j is a tautology, this can only be because each A'_i and B'_j share some atom. Thus $A' \rightarrow B'$ is a tautological entailment, and so it is easy to see that its substitution instance $A \rightarrow B$ is a tautological entailment.

We have just proven

THEOREM 1. *$A \rightarrow B$ is a first-degree entailment of Anderson and Belnap's system E iff A entails B according to the L Criterion filled out according to suggestion 1.*

We turn now to discussing suggestion 2 as to how to define the key notions in the L Criterion of "partial tautology" and "partial contradiction." Suggestion 2 uses no notion of "equivalence" whatsoever, and so, *a fortiori*, no such notion that is open to the charge of being *ad hoc*. Instead, the basic apparatus of suggestion 2 is the semantically very natural notion due to Smullyan of analytic tableaux.

In JEFFREY (1967), which makes an excellent elementary introduction to analytic tableaux, they are called 'truth trees,' and I shall here use this suggestive name. The basic idea of a truth tree is that the truth conditions for a given sentence are diagrammed in a branching way, each branch representing a possible way in which the sentence could be true. This is not quite right; some branches may represent "impossible ways" in which the sentence "could" be true. These are the *closed* branches which contain

some sentence and its negation. The rules for truth tree construction are:

$$(T\&) \frac{A \& B}{\begin{array}{c} A \\ B \end{array}} \quad (T\neg\&) \frac{\sim(A \& B)}{\begin{array}{c} \sim A \\ \sim B \end{array}} \quad (T\vee) \frac{A \vee B}{\begin{array}{c} A \\ B \end{array}} \quad (T\neg\vee) \frac{\sim(A \vee B)}{\begin{array}{c} \sim A \\ \sim B \end{array}}$$

$$(T\neg\neg) \frac{\sim\sim A}{A}$$

The perfectly dual notion of a “falsity tree” is for some reason not so familiar. A falsity tree diagrams in a branching way what sentences must be false in order for a given sentence to be false. The rules for falsity tree construction are:

$$(F\&) \frac{A \& B}{\begin{array}{c} A \\ B \end{array}} \quad (F\neg\&) \frac{\sim(A \& B)}{\begin{array}{c} \sim A \\ \sim B \end{array}} \quad (F\vee) \frac{A \vee B}{\begin{array}{c} A \\ B \end{array}} \quad (F\neg\vee) \frac{\sim(A \vee B)}{\begin{array}{c} \sim A \\ \sim B \end{array}}$$

$$(F\neg\neg) \frac{\sim\sim A}{A}$$

It is well known that a sentence is contradictory iff it has a truth tree in which all branches are closed (such a tree is itself called *closed*—otherwise *open*). And there is of course the perfectly dual fact that a sentence is a tautology iff it has a closed falsity tree.

It is natural then to define a sentence as a *partial contradiction* [*partial tautology*] iff it has a truth [falsity] tree with some closed branch (call such a tree *partly closed*—otherwise *fully open*). The intuitive idea is to require of a partial contradiction [partial tautology] that at least one of its truth [falsity] conditions be reducible to absurdity.

Using these definitions, the preferred result of the paper may be obtained:

THEOREM 2. *A → B is a first-degree entailment of Anderson and Belnap's system E iff A entails B according to the L Criterion filled out according to suggestion 2.*

I shall not here attempt a detailed proof of Theorem 2 (such a proof can be found in DUNN, 1980), but I would like to make clear the connection of the L Criterion spelled out using truth (falsity trees) and the Anderson–Belnap normal form sharing test of tautological entailments. The basic idea is that a truth tree can be thought of as working out the disjunctive normal form of a sentence (conjoin the atoms in each branch,

and then disjoin all these conjunctions). Dually a falsity tree is connected to conjunctive normal forms. It is easy to connect the rules of tree construction with normal form rules and thereby produce a proof of Theorem 2 from Theorem 1 (a different proof is found in DUNN, 1980).

References

- ANDERSON, A. R., and N. D. BELNAP, Jr., 1962, *The pure calculus of entailment*, The Journal of Symbolic Logic, vol. 27, pp. 19–52
ANDERSON, A. R., and N. D. BELNAP, Jr., 1975, *Entailment*, vol. 1 (Princeton University Press, Princeton)
DUNN, J. M., 1980, *A sieve for entailments*, The Journal of Philosophical Logic, vol. 9, pp. 41–57
JEFFREY, R. C., 1967, *Formal logic: Its Scope and limitations* (New York)
LEWY, C., 1958, *Entailment*, Aristotelean Society Supplementary, vol. 32, pp. 123–142
LEWY, C., 1976, *Meaning and modality* (Cambridge University Press, Cambridge, England)

TRUTH-VALUE GAPS*

JOHN McDOWELL

University College, Oxford, England

FREGE (e.g. 1892, pp. 32–33) and others (e.g. STRAWSON, 1950) have held that if one utters an atomic sentence containing a singular term which lacks a denotation, then one expresses neither a truth nor a falsehood. I want to contrast two justifications for that thesis.

I

According to DUMMETT (1958–9, 1960, 1973, Chapter 12, 1978, pp. xiv–xviii), the only justification lies in the smoothness which the thesis permits, in an account of how atomic sentences function as constituents of complex sentences.

The background is a distinction (DUMMETT, 1973, p. 417) between two ways of approaching the notion of truth-value, in the context of the idea that a theory of meaning for a language might centre on the notion of truth. In the first approach, the notion of truth-value constitutes the point of connection between, on the one hand, an account of what it is to make an assertion, and, on the other, the general form of statement whereby the theory determines the content of assertions which can be effected by uttering sentences, simple or complex, in the language. In the second approach, truth-values are ascribed to sentential constituents of complex sentences, in such a way as to facilitate a systematic account of their impact on the truth-values of the sentences of which they are constituents.

Of course, these two approaches cannot be wholly disconnected. For the systematic account of sentential compounding which the second approach would yield could have no point other than to subserve the needs of

* This paper owes a great deal to conversations over many years with Gareth Evans.

a systematic determination of the content of assertions effected by uttering whole (complex) sentences. So its assignments of truth-values to whole (complex) sentences would have to conform to whatever requirements the first approach imposes. In particular, they would have to respect the thought that the notion of truth which the first approach needs is anchored in the grasp we acquire, in learning to speak, of what it is for an assertoric utterance to be correct. However, there is no *a priori* assurance that the way in which the notion of truth-value is employed in the second approach, in connection with a sentence considered as a potential component of complex sentences, will correspond neatly with the way in which it is employed, in connection with the very same sentence as used on its own to make an assertion, in the first approach.

Dummett's justification, now, appeals to considerations from the second approach. Specifically: it is natural to take a sentence of the form "*a* is not *F*" as the negation of the corresponding sentence of the form "*a* is *F*"; and it is natural to connect negation and falsehood, by way of the principle that a sentence is false if and only if its negation is true. If we were to count a sentence of the form "*a* is *F*" false when the singular term lacks a denotation, then the two natural thoughts would commit us to counting the corresponding sentence of the form "*a* is not *F*" true; but that would not cohere with the indispensable connection between the notions of truth and correctness in assertion. (Here the second approach is, as noted, constrained by the first.) Obviously, counting the original sentence true would directly flout the indispensable connection. So if we find it desirable to preserve the two natural thoughts, that constitutes a justification for counting such sentences neither true nor false.

These considerations do not preclude Dummett from insisting, as he does, that employment of the notion of truth-value in the first approach must conform to the principle *tertium non datur*, on the ground that (vagueness and ambiguity aside) our understanding of what it is to make an assertion leaves no room for a gap between the conditions under which an assertion is correct and the conditions under which it is incorrect. This argument applies in particular to the utterance, with assertoric intent, of a sentence of the form "*a* is *F*". According to Dummett's argument, the sense of such an utterance is determined, like that of any assertoric utterance, by the distinction between the condition under which it would be correct and all other conditions. The content of an assertion effected by such an utterance is such as to rule out any condition whose obtaining would render it incorrect. Conditions thus ruled out include both that in which our consider-

ations from the second approach permit us to count the sentence false, and that in which those considerations recommend counting the sentence neither true nor false. Thus if we use the word "false" in the way that seems most natural from the standpoint of the first approach, the states which the second approach led us to distinguish, as being false on the one hand and being neither true nor false on the other, appear rather as two different ways of being false.

A terminological manoeuvre will remove the confusing appearance that the two approaches conflict. The simplest proposal is probably this: to concede the word "false" to the second approach, but to defer to the requirement of *tertium non datur* in the first, by collecting the states described as "being false" and "being neither true nor false" together as possession of undesignated truth-values, as opposed to the designated truth-value, truth. The principle *tertium non datur* now takes this form: vagueness and ambiguity aside, no assertion has either a designated nor an undesignated truth-value.

On this view, then, the line of thought which justifies Frege's "neither true nor false" thesis would culminate in the following idea: if singular terms without denotations can occur in a language, then the impact of negation on the truth-values of negative sentences is best captured, not by the usual two-valued truth-table, but by a three-valued table, with the additional stipulation that negating a sentence which has the undesignated truth-value called "being neither true nor false" yields a sentence with that same truth-value. A full elaboration of the idea would require, also, that we specify the impact of a constituent with this third truth-value on the truth-values of other sorts of complex sentence, by providing three-valued truth-tables for the other sentential connectives as well.

Note that the third truth-value is just that, a truth-value. Strictly speaking, on this view, there is no question of a truth-value gap. This means that, about Frege himself, we can suppose at most that he took only the first step along the path which Dummett describes; for what Frege held was that the sentences in question lacked truth-values altogether. According to Dummett, what this stemmed from—the obstacle which debarred Frege from the natural culmination of his intuition—was the doctrine that sentences stand to their *Bedeutungen*, truth-values, in a relation which is not just analogous to, but a case of, the name-bearer relation.¹ For it is a com-

¹ DUMMETT (1973), pp. 185–186, 427–429. I suppress for simplicity the ambiguity which Dummett finds in the notion of *Bedeutung*.

pelling principle that if a complex name has a constituent which lacks a *Bedeutung*, then the complex name lacks a *Bedeutung* too. If Frege had allowed himself to suppose that the relation between a sentence and its truth-value was only analogous to the relation between a complex name and its bearer, then it need not have seemed obvious that the principle must apply to the case at hand, and Frege would not have been precluded from the three-valued theorizing which lies at the end of the path on which, according to Dummett, he started. Taking the principle to apply, however, Frege thought natural language confronted him with the possibility that a sentence might have a sense but (since it lacked a *Bedeutung*) lack a semantic role; and, given that view of the situation, it was reasonable for him to suppose natural language beyond the reach of coherent theory.

In the three-valued treatment, the conditions for the truth and the falsity of a sentence of the form “*a* is *F*” overlap: both include the absence of the condition under which the sentence is said to be neither true nor false. We might introduce a label, say “presupposition”, for the relation between the sentence, or an utterance of it, and this overlapping condition. (DUMMETT, 1978, p. xiv.)

Thus introduced, the notion of presupposition is clearly not fundamental in the way that the notion of assertion is. That a certain utterance presupposes that a certain condition obtains is not conceived as an hypothesis which casual observation of the practice of speaking a language might recommend to us, independently of theorizing about the internal structure of sentences, in the same sort of way as it might recommend the hypothesis that a certain utterance is an assertion that a certain condition obtains; as if it might be independently attractive to preserve hypotheses of both sorts, if possible, in the subsequent task of theorizing about how the structure of utterances bears on their correct interpretation. On the contrary: this notion of presupposition emerges only in the course of theorizing about structure. The aim of the theorizing is to secure the neatest possible fit between the theory’s deliverances about the correctness and incorrectness of assertions, on the one hand, and the observable practice of speaking the language, on the other; and whereas the concept of assertion occupies a point of direct contact between an acceptable theory of a language and what its speakers observably do, the concept of presupposition has its utility, if any, only inside the theory.

Dummett contrasts this view of presupposition with a view according to which presupposition is as fundamental as assertion; so that the use of the notion is intelligible, and potentially informative about the meanings

of utterances, without benefit of information or theory about structure. Suppose we have a pair of sentences, *A* and *B*, each of which can be correctly asserted if and only if both of a pair of conditions, *C* and *C'*, obtain. Then this different view of presupposition would involve the idea that, independently of any account of structure, we can be told something intelligible about a difference in meaning by being told that in the case of *A*, but not *B*, the obtaining of *C* is only a presupposition. As DUMMETT insists (1978, pp. xv–xviii), this idea is utterly implausible.

Frege himself remarks that the assertoric use of a sentence containing a singular term presupposes that the term designates something; and he does not embark on the three-valued theory which alone, according to Dummett, gives the notion of presupposition its proper theoretical context. But this does not license accusing Frege of treating presupposition as independently fundamental. Frege's remark about presupposition expresses the same intuition about natural language as the "neither true nor false" thesis: an intuition which seems to Frege to preclude systematic theory about natural language as it stands. The remark about presupposition is not a fragment of a serious theory; rather, that it is needed would strike Frege as a reflection of the very fact about natural language which makes serious theory impossible.

STRAWSON puts the notion of presupposition to a similar use, equally without a backing of three-valued truth-tables (1950 (with different terminology), 1952, 1964). And Strawson lacks Frege's reason for supposing that natural language is not amenable to serious theory; indeed, he evidently takes the notion of presupposition to be an essential ingredient in any serious account of referring in natural languages. Since Dummett thinks it is only in the context of three-valued truth-tables that the notion of presupposition, if regarded as theoretically important at all, has its properly secondary role, he can thus interpret Strawson's failure to concern himself with three-valued truth-tables only as an indication that Strawson holds the implausible view distinguished above: namely that the notion of presupposition is on a level with the notion of assertion (DUMMETT, 1978, p. xviii).

II

A different justification, which Dummett refuses to countenance, turns on the following idea. The syntax of sentences of the relevant sort fits them to express singular thoughts if any; where a singular thought is

a thought which would not be available to be thought or expressed if the relevant object, or objects, did not exist. It follows that if one utters a sentence of the relevant sort, containing a singular term which, in that utterance, lacks a denotation, then one expresses no thought at all; consequently, neither a truth nor a falsehood.

This conception of singular thoughts, which is in essence RUSSELL'S (1956)² must be separated from two accretions which would preclude using it for the present purpose.

First, Russell held that, if members of a syntactic category of apparent singular terms can lack denotations, then even those members of the category which have denotations are not genuine singular terms. For Russell, lack of denotation on the part of a putative singular term shows, not that utterances of sentences in which it occurs express no thoughts, but that utterances of all sentences in the relevant syntactic category express non-singular thoughts. This is the line of argument which convinces Russell that the only genuine singular terms are logically proper names. The conception of singular thoughts characterized above promises to open truth-value gaps, but this line of argument would ensure that the promise is not fulfilled.

Second, Russell found it natural to describe singular thoughts (or propositions) as propositions in which objects themselves occur.³ It may seem an obvious gloss on this to say that a Russellian singular thought, of a sort expressible by using a one-place predicate, is well represented as an ordered couple whose members are the appropriate object and (perhaps) the appropriate property.⁴ In that case there cannot be two different singular thoughts which ascribe the same property to the same thing. It follows that a genuine singular term—one fitted for the expression of singular thoughts—is one which allows no room for a distinction between sense and reference. If this were a necessary corollary of the Russellian conception of singular thoughts, then that conception could not figure in an account of a possibly Fregean ground for the “neither true nor false” thesis.

² pp. 45–47 (though this passage also contains the argument for the first accretion applied to definite descriptions).

³ RUSSELL (1956), p. 45: the view of “denoting phrases” against which he is arguing represents them “as standing for genuine constituents of the propositions in whose verbal expressions they occur”.

⁴ This obviously generalizes to relational thoughts. For the ordered-couple conception of monadic singular thoughts, see, e.g., DONNELLAN (1974), pp. 11–12.

Both these accretions are detachable.

As for the first: Russell achieves his radical restriction of genuine singular terms by applying, to any category of apparent singular terms whose members can lack denotations, an argument which runs in effect as follows.⁵

- (1) If such terms are genuine singular terms, then one expresses no thought by uttering an atomic sentence containing a denotationless member of the category.⁶
- (2) One does express thoughts by such utterances.⁷

Therefore

- (3) Such terms are not genuine singular terms.

(1) formulates the conception of singular thoughts from which I aim to detach the accretion. But the accretion requires (2) as well, for any category of atomic sentences utterances of members of which one can take oneself to understand even if one is mistaken in supposing that a suitable object exists. The upshot is, in effect, that one can understand an utterance as expressing a singular thought only if one's conviction that there is an object for the thought to be about is proof against Cartesian doubt. If, where that is not so, we continue to suppose that we understand some of the utterances as expressing singular thoughts, then we are committed to the idea that the impression of understanding, in the other cases, is an illusion: one takes oneself to understand an utterance as expressing a singular thought, but the singular thought which one thinks one understands the utterance to express does not exist. (2) disallows this: it registers insistence that the impression of understanding cannot be an illusion. But what grounds are there for this insistence?

⁵ Reconstructed from RUSSELL (1956), pp. 45–47—the application to definite descriptions.

⁶ RUSSELL says (1956, p. 45) that the sentence “ought to be nonsense”. As against Frege (on the standard interpretation: but see below)—the target of Russell's argument—this is inept; for Frege (on that interpretation) is at pains to equip such terms, and hence sentences containing them, with sense. But what Russell is expressing is an inability (with which we can sympathize: see below, on the intuition which I seek to detach from the second accretion) to see how a term can both be a genuinely singular term and have a sense indifferent to the non-existence of anything for it to refer to. (It is not clear, *pace* the standard interpretation, that Frege thought these could be combined either: see the end of this paper.)

⁷ Russell says that the sentence “is plainly false”. I take this to express the conviction that the sentence expresses a thought, together with a resolve to use the word “false” in the way which would be appropriate in the first of the two approaches to the notion of truth-value distinguished in part I.

One might think one could support Russell's argument on the following lines. If an utterance expresses a singular thought, it must be by virtue of its logical form that it does so. Thus, when we contemplate resisting the argument, as applied to some category of utterances, by exploiting the idea of an illusion of understanding, we are contemplating treating the category as having some members which have the appropriate logical form, and others which do not. The difference between the two sub-categories would turn on facts about what does and what does not exist. Ignorance of such facts is not the sort of thing we usually take to impugn someone's competence in a language. But should it not impugn someone's competence in a language if he takes an utterance to have a logical form which it does not have? (DUMMETT, 1973, p. 163.)

The basis of this defence is the following idea. Granting that logical form may diverge from superficial syntactic form, nevertheless superficial syntactic form is all that is presented to a hearer of an utterance. Hence if understanding an utterance—which involves sensitivity to its logical form—is an exercise of competence in a language, then logical form should be recoverable from superficial syntactic form by “pure” linguistic knowledge, without one's needing to draw also on “extra-linguistic” knowledge, such as would be constituted by knowledge of what does and what does not exist.

But this is inconclusive. I am not contemplating a possibility in which logical form is cut completely adrift from syntactic form. That would indeed make a mystery of the connection between syntax and the capacity to understand utterances in a language. But resistance to Russell's argument, on the lines I am contemplating, allows that syntactic form determines logical form, to this extent: it is recoverable from the syntactic form of an utterance of one of the relevant kinds, without knowledge of what does and what does not exist, that the utterance has subject-predicate logical form if any; that is, that it expresses a singular thought if any. Here “if any” leaves a question open, to be resolved by knowledge of whether or not an appropriate object exists. If one thinks one can object to this suggestion by claiming that this knowledge is extra-linguistic, one must suppose that “pure” linguistic competence, conceived as untainted by knowledge of what does and what does not exist, ought to carry its possessor all the way to the thought expressed by any utterance he understands. But that supposition is not an independently obvious truth, something one could properly appeal to in order to eliminate resistance to Russell's argument. On the contrary, it is at least as compelling to contrapose:

if a language can encompass the expression of singular thoughts about the sorts of items our knowledge of whose existence is a substantial and precarious achievement, then understanding some utterances must involve bringing to bear knowledge of just that kind—with the attendant risk that the appearance of understanding may be illusory.⁸ If the knowledge is extra-linguistic, then so much the worse for the idea that “pure” linguistic competence suffices for the understanding of absolutely any utterance in a language.

(Whether knowledge of existence is plausibly thought of as extra-linguistic may, anyway, vary from case to case. Falsely supposing that there is a denotation for some utterance of “that lime-tree” need not impugn one’s linguistic competence; whereas the belief that “Vulcan” had a bearer was arguably a defect in command of the language.⁹)

As for the second accretion: it is true that the ordered-couple conception of singular thoughts would yield the thesis that singular thoughts depend for their existence on the existence of the objects they are about. But it is not true that one can embrace the thesis only by endorsing the ordered-couple conception. On the contrary, there is a plausible way of formulating the thesis without commitment to that conception. This alternative exploits the idea that in order to specify a thought-content, which one does typically in a “that” clause, one must express the thought oneself.¹⁰ In this context, a singular thought is a thought which one cannot ascribe to someone, or assign as the content of an assertion, without oneself making a reference to the appropriate object. Now if one knows that the existence of an appropriate object is an illusion, then one cannot specify a thought-content in the appropriate form. There is, one knows, no thought which one could ascribe in that form. This is the Russellian thesis—formulated here in terms of reflections about “that” clauses which are compatible with denying what the ordered-couple conception implies, namely that co-referring singular

⁸ Cf. FREGE (1918/19), p. 73: “By the step with which I secure an environment for myself I expose myself to the risk of error”. By “error” Frege here means “lapsing into fiction”; on which see the last paragraph of this paper.

⁹ The envisaged illusion of understanding, in the case of, e. g., an utterance of a sentence like “That lime-tree is covered with leaves”, need not be an illusion of understanding the *sentence*. Russell’s argument would work, for cases of this kind, if it followed from the meaningfulness of a sentence that any utterance of it expressed a thought. But it does not follow: see STRAWSON (1950).

¹⁰ This idea is well captured by, though it does not require, the view of “that” clauses suggested by DAVIDSON (1969).

terms are interchangeable *salva veritate* in "that" clauses of the relevant sort (cf. McDowell, 1977).

As standardly interpreted, Frege's distinction between sense and reference of singular terms has two characteristic elements: first, that such terms can possess the sort of sense appropriate to them whether or not they refer to anything; and, second, that two such terms can differ in sense (hence, fail to be interchangeable *salva veritate* in thought-ascribing "that" clauses) while referring to the same thing. If we construe the thesis that objects occur in singular propositions as an expression of the ordered-couple conception, then we are taking it as a blanket rejection, on Russell's part, of both elements of that standard interpretation. Now Russell certainly found no merit in the doctrine of sense and reference. But it is possible, and arguably charitable, to detach the thesis from the main body of Russell's attacks on the doctrine, and construe it as expressing a laudable recoil from the first element of the standard interpretation, without in itself involving commitment as to the second.

On this view, the thesis expresses an intuition on the following lines. If a term has a sense which is indifferent to the non-existence of any suitably related object, then it is not recognizable as a singular term—one whose sense fits utterances containing it to express thoughts which are about an object, on a certain attractive conception of what that amounts to. If an object's non-existence would not matter for the existence of certain thoughts, then the object's relation to those thoughts falls short of an intimacy which, Russell insists, sometimes characterizes the relation of things to thoughts, namely that the thoughts would not exist if the things did not. The difference between thoughts which have this intimate relation to objects and thoughts which do not is sufficiently striking to deserve to be marked by the stipulation that only the former should count as being in the strictest sense about objects.

This intuition has no necessary connection with the ordered-couple conception. The intuition recognizes a relation between objects and thoughts so intimate that it is natural to say that the objects figure in the thoughts; but it has no tendency to recommend the idea that, for a given object which can thus figure in our thoughts, there is only one way in which it can figure—only one mode of presentation. (See, further, McDowell, 1977; Evans, 1980.)

It seems appropriate to credit this intuition to a robust sense of reality. If we insist that we think some thoughts of Russell's singular kind, then we conceive objects as sometimes present to our thoughts, in a way which con-

trasts with the most that is available on a picture according to which we could have all the thoughts we are entitled to think we have even if no objects besides ourselves existed. It is a pity that Russell allowed the first accretion to override his robust sense of reality, and convince him that the second picture was the best that could be had, in all cases in which Cartesian epistemology would represent the conviction that an appropriate object exists as uncertain. And if we let the second accretion persuade us to abandon our robust sense of reality wherever the ordered-couple conception is inappropriate (because it matters how one refers to an object in the specification of a thought), then we arrive at a similar, and similarly deplorable, upshot: like Russell in consequence of the first accretion, we almost lose hold of the best of Russell's insights into the relation between thought and things.

If the first element is anyway a misinterpretation of Frege, then the intuition is even less anti-Fregean than I have so far suggested. I shall return to this at the end.

Three-valued truth-tables seem necessary only on the assumption that an utterance of one of the problematic sentences, with assertoric intent, constitutes a "move in the language-game"; so that the sentence must be credited with a semantic role—hence, a *Bedeutung*. But according to the second justification, an utterance of the problematic kind, though it may masquerade as a "move in the language-game" of the kind constituted by the assertoric expression of a singular thought, in fact simply fails to be what it purports to be. So on this view there are genuine truth-value gaps.

The notion of presupposition has a natural use in this position, for the relation which a singular sentence, or an utterance of it, bears to the condition, or conditions, which must be satisfied if the utterance is to express a thought. This use of the notion of presupposition emerges from, and is intelligible only in the context of, reflections about how the structure of singular sentences suits them to express the kind of thought they are capable of expressing. It is thus unfair to claim, as Dummett does, that anyone who makes serious use of the notion without a backing of three-valued truth-tables must take the notion to be "given naturally and in advance of the analysis of any particular forms of sentence" (DUMMETT, 1978, p. xviii).¹¹

¹¹ DUMMETT complains (1973, p. 423) that "in the absence of a distinction between designated and undesignated truth-values, the mere proposal to regard a certain kind of sentence as being, in certain kinds of case, neither true nor false does not tell us whether the state of being neither true nor false is to be regarded as a sub-case of correct assertibility or of

What Dummett overlooks is the possibility of locating the use of the notion in the context of considerations about the structure, not of complex sentences, but of atomic sentences themselves. (Dummett appeals to facts about how it is natural to use the word "false" in theorizing about sentential compounding, specifically negation; within the second position, these facts will seem pointers to the correctness of what the position claims about the structure of the atomic sentences which figure in such compounding.)

Dummett himself sees no merit in the idea which underlies this second justification for the "neither true nor false" thesis. He refuses to allow that there can be the sort of illusion of understanding discussed above.¹² Thus he endorses, in effect, the leading idea of the first accretion; and, since he has no truck with Russell's conception of the logically proper name, this means that he makes nothing of the intuition which I credited to a robust sense of reality. But this refusal to allow illusions of understanding, so far from being, as Dummett suggests, the only alternative to an obvious absurdity, requires a view which, from the standpoint of the robust sense of reality, looks quite unattractive: a view according to which knowing one's way about in a language—being able to recognize the thoughts expressed in it—is prior to and independent of knowing one's way about in the world.¹³

It may seem that Dummett is on stronger ground when he refuses to read the second justification into Frege. The first element of the standard interpretation would certainly preclude this. But it is not obvious that that element is correct. The question is too complex to discuss properly here; but I shall mention two difficulties for the standard interpretation.

incorrect assertibility, and hence does not determine the assertoric content of the sentence". But in the present case the proposal does not purport to effect such a determination; it flows from such a determination (namely that, in the relevant circumstances, the sentence has no assertoric content), independently effected on the ground of the sentence's internal structure.

¹² (1973), p. 404: the idea that one of the problematic utterances expresses no thought "would be absurd, since we can understand such an utterance, and if we wrongly suppose the name to have a bearer, we can also believe it". But that we can really understand the problematic utterances is not an argument against the second justification, but rather exactly what such an argument would need to make out. And, of course, it does not follow, from the fact that we may believe that such an utterance expresses a truth (which is the most that is uncontroversial), that we may believe some proposition expressed by such an utterance.

¹³ Dummett also adduces, as a reason against truth-value gaps, the implausibility of the presumed consequence that some truth-functional compounds would lack truth-value (DUMMETT, 1973, pp. 425–426). But see GEACH (1976), p. 441.

First: Dummett's view certainly entitles us to introduce a notion of presupposition. But it does not entitle us to distinguish what an utterance presupposes from the content of an assertion effected by it.¹⁴ Precisely not: on Dummett's view the content of an assertion is fixed by the line between the condition under which the assertoric utterance of a sentence is correct and all other conditions; the content is such as to rule out all these other conditions—including a presupposition's failing. There is thus no explanation here for a view according to which the fulfilment of presuppositions is not part of what is asserted when a sentence is assertorically uttered. But this view is perfectly intelligible in the context of the second justification for the “neither true nor false” thesis. And it is evidently this view, according to which presuppositions are distinguished from content, rather than a distinguished component within content, to which Frege is attracted.¹⁵

Second: in those passages which constitute the evidence for the first element of the standard interpretation, there is typically an appeal to the notion of fiction.¹⁶ Frege's use of the notion of fiction is peculiar: he cites examples of fictionally intended utterances, but he also uses the notion in such a way that it is possible to lapse into fiction without knowing it—this is what happens, in his view, whenever one utters, with serious assertoric intent, a sentence containing a denotationless singular term. Now the idea that one can unknowingly lapse into fiction is so wrong-headed about fiction that we urgently need an account of why it should have attracted so penetrating a thinker.¹⁷ A satisfying explanation is suggested by a revealing passage in the posthumously published *Logik* of 1897 (FREGE, 1969), in which Frege writes that in fiction we are concerned with apparent thoughts and apparent assertions, as opposed to genuine thoughts, which are always either true or false. This coheres neatly with the idea that

¹⁴ Though Dummett unaccountably claims this, at DUMMETT (1978), p. xiv.

¹⁵ FREGE (1892), p. 40, implicitly denies “that the sense of the sentence ‘Kepler died in misery’ contains the thought [its presupposition] that the name ‘Kepler’ designates something”. The context makes it difficult to argue that the topic is ingredient sense, as opposed to content.

¹⁶ E. g. FREGE (1892), pp. 32–33. I owe to Gareth Evans my appreciation of the peculiarity of Frege's use of the notion of fiction, and of the importance of this for understanding his views about reference.

¹⁷ Cf. DUMMETT (1973), p. 160: “We should not, as Frege often does, cite as examples of names having sense but no reference personal names used in fiction... We need names used with a serious, though unsuccessful, intention to refer”. What Dummett misses is that for Frege any case of a name used with a serious, though unsuccessful, intention to refer is a case of fiction.

Frege's ground for the "neither true nor false" thesis was on the lines of the second justification. It suggests that what attracted Frege to his peculiar use of the notion of fiction was that it seemed to soften the blow of the implication that there is an illusion of understanding. By the appeal to fiction, Frege equips himself to say that it is not a complete illusion that one understands one of the problematic utterances, any more than it is an illusion that one understands an overtly fictional utterance. But the 1897 passage shows that in Frege's view the understanding of the problematic utterances which he entitles himself to recognize is separated by a great gulf from understanding of informative speech. It would be in the spirit of his talk of apparent thoughts to talk of apparent understanding; certainly the belief that one understands one of the problematic utterances as expressing a genuine thought would be an illusion, just as the second justification for the "neither true nor false" thesis requires. If this is the point of Frege's appeal to fiction, then the standard passages do not undermine the attribution to him of the second justification, or support the ascription to him of the thesis with which he is usually saddled: that a genuine singular term—one suited for use in the expression of genuine thoughts, as opposed to a *Scheineigenname*—has a sense (an impact on the thoughts expressible by sentences containing it) which it could have whether or not it referred to anything.¹⁸

References

- DAVIDSON, D., 1969, *On saying that*, in: Words and objections, eds. D. Davidson and J. Hintikka (Reidel, Dordrecht), pp. 158–174
 DONNELLAN, K. S., 1974, *Speaking of nothing*, Philosophical Review, vol. 83, pp. 3–31
 DUMMETT, M., 1958–9, *Truth*, Proceedings of the Aristotelian Society, vol. 59, pp. 141–162
 DUMMETT, M., 1960, *Presupposition*, The Journal of Symbolic Logic, vol. 25, pp. 336–339
 DUMMETT, M., 1973, *Frege: Philosophy of language* (Duckworth, London)

¹⁸ Although Evans does not saddle Frege with the thesis as a general claim about singular terms (see especially EVANS, 1980), he does ascribe to Frege the thesis that there is a category of genuine singular terms (called "Fregean" in EVANS, 1979) whose sense is existence-indifferent. Russell would have seen no naturalness in a classification which grouped these together with "Russellian" singular terms, but EVANS aims (1979) to show that there is a semantic natural kind which includes both. I do not want to express a view on the success of this endeavour. But I believe the considerations in the text disarm the passages Evans might cite to justify counting Frege as an ally. I know no clear evidence that Frege believed in genuine singular terms—as opposed to *Scheineigenamen*—which were "Fregean"; I think we can credit Frege with a view of the natural boundaries which matches the core of Russell's view (and has the enormous advantage of avoiding the two accretions).

- DUMMETT, M., 1978, *Truth and other enigmas* (Duckworth, London)
- EVANS, G., 1979, *Reference and contingency*, The Monist, vol. 63
- EVANS, G., 1980, *Understanding demonstratives*, in: Meaning and understanding, eds. J. Bouveresse and H. Parret (De Gruyter, Berlin)
- FREGE, G., 1892, *Ueber Sinn und Bedeutung*, Zeitschrift für Philosophie und philosophische Kritik, vol. 100, pp. 25–50 (Quoted from: Translations from the Philosophical Writings of Gottlob Frege, eds. P. Geach and M. Black (Blackwell, Oxford, 1952))
- FREGE, G., 1918/19, *Der Gedanke*, Beiträge zur Philosophie des deutschen Idealismus, vol. 1, pp. 58–77 (Quoted in the translation of A. M. and M. Quinton, Mind 65 (1956), pp. 289–311)
- FREGE, G., 1969, *Nachgelassene Schriften*, eds. H. Hermes, F. Kambartel, F. Kaulbach (Meiner, Hamburg)
- GEACH, P. T., 1976, *Critical notice of Dummett*, [1973], Mind, vol. 85, pp. 436–449
- MCDOWELL, J., 1977, *On the sense and reference of a proper name*, Mind, vol. 86, pp. 159–185
- RUSSELL, B., 1956, *Logic and knowledge*, ed. R. C. Marsh (Allen and Unwin, London)
- STRAWSON, P. F., 1950, *On referring*, Mind, vol. 59, pp. 320–344
- STRAWSON, P. F., 1952, *Introduction to logical theory* (Methuen, London)
- STRAWSON, P. F., 1964, *Identifying reference and truth-values*, Theoria, vol. 30

SEMANTICS OF GENERALIZED STATE DESCRIPTIONS

E. K. VOISHVILLO

The main purpose of this paper is to analyze the relation of entailment in the intensional (relevant) sense (as the intensional connection between propositions) for some propositional logical systems. Classical and S5 systems will be considered here. As it has been noticed by numerous authors (ROUTLEY and ROUTLEY, 1972, DUNN 1976, ANDERSON and BELNAP, 1975, § 16.2.1) for defining the indicated relations in the intensional sense (without paradoxes of the type $A \& \neg A \models B$ and $B \models A \vee \neg A$), it is necessary to assume that in some sense $A \& \neg A$ may be true and $A \vee \neg A$ may not be true. It means that using the notion of state description (s.d. in short) to analyze the relation of entailment, e.g., for propositions in the language of classical propositional logic, one should allow inconsistent and incomplete state descriptions regarding some propositional variables though they seem unnatural as descriptions of possible states of affairs.

We want to show how one can come to this generalization of the classical s.d. notion in natural way by analyzing the sources of "paradoxes" in classical logic and to build up for it a respective semantics of entailment. Besides, the possibility of generalization of the same s.d. notion in another direction will be shown, namely, introducing s.d. with certain dependencies between their elements. The informal semantics for the system S5 is built up on the basis of this kind of notions. The possibility of the entailment definition for the system in the intensional sense is shown.

Explication of entailment as an information relation between propositions (determined by logical forms of propositions) is primary for all systems. Just this meta-assertion " $A_0 \models B_0$ " (" B_0 is logical consequence of A_0) is understood as "The information of B'_0 is a part of the information of A'_0 ",

for every propositions A'_0 and B'_0 having the same logical forms as A_0 and B_0 . $A_0 \models B_0$ is equivalent to $A \models B$.

Bearing in mind the qualitative notion of information we define information of propositional form A (logical form of A'_0) as a pair $\langle M_A, M \rangle$ where M is an s.d. set and M_A is its subset where A is true. This corresponds to the interpretation (accepted in the semantic theory of information) of the information of A as a measure of restriction of M conditioned by acceptance of A . The information of $A'_0 - I(A'_0)$ may be interpreted as $I(A/\Gamma)$ (information of A in the presence of Γ data) and defined as $\langle (M_\Gamma)_A, M_\Gamma \rangle$ where Γ is a set of laws of some theory associated with the interpretation of descriptive terms of the language transforming A into A'_0 . Γ forbids some s.d., restricting M to M_Γ . The relation " $I(B'_0)$ is a part of $I(A'_0)$ " is naturally defined as " $(M_\Gamma)_A \subseteq (M_\Gamma)_B$ ". This leads in combination with the accepted definition of " \models " to the correlation

$$(1) A \models B \Leftrightarrow \dot{\forall} \Gamma ((M_\Gamma)_A \subseteq (M_\Gamma)_B).$$

In particular, for classical logic:

$$(2) A \models B \Leftrightarrow M_A \subseteq M_B \Leftrightarrow I(B) \text{ is a part of } I(A).$$

Classical propositional logic (C. P. L.)

The correlation (2) presents a well-known definition of classical logic entailment if M is a set of classical s.d., i.e., conjunctions of the type $(\tilde{p}_1 \& \tilde{p}_2 \& \dots \& \tilde{p}_n)$ where p_1, p_2, \dots, p_n are different in pairs propositional variables (logical forms of elementary propositions); \tilde{p}_i ($i = 1, 2, \dots, n$) is p_i or $\neg p_i$. It will be more convenient for us to understand s.d. as a set $\{\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_n\}$.

To define M_A let us write out well-known truth and falsity conditions of the formula A in s.d. (" TA/α " stands for " A is true in α " and " FA/α "—" A is false in α "). We imply a language with the following connectives: $\&$, \vee , \neg (conjunction, disjunction, negation).

$$\text{DF 1. } Tp_i/\alpha \Leftrightarrow p_i \in \alpha; Fp_i/\alpha \Leftrightarrow \neg p_i \in \alpha \quad (i = 1, 2, \dots, n, \dots).$$

$$T(A \& B)/\alpha \Leftrightarrow TA/\alpha \text{ and } TB/\alpha; F(A \& B)/\alpha \Leftrightarrow FA/\alpha \text{ or } FB/\alpha;$$

$$T(A \vee B)/\alpha \Leftrightarrow TA/\alpha \text{ or } TB/\alpha; F(A \vee B)/\alpha \Leftrightarrow FA/\alpha \text{ and } FB/\alpha;$$

$$T\neg A/\alpha \Leftrightarrow FA/\alpha; F\neg A/\alpha \Leftrightarrow TA/\alpha.$$

$$\text{DF 2. } M_A = \{\alpha: TA/\alpha\}.$$

The above definition of " $A \models B$ " as " $I(B)$ is a part of $I(A)$ " (Correlation (2)) permits to find out the reason for paradoxicality of the classical

notion of entailment. The point is that here we imply $I(A/\Gamma)$ instead of $I(A)$ —the information which is added by A given a set of some data Γ_0 where Γ_0 in this case is the set of conditions for s.d.:

- (a) $\dot{\forall}\alpha\dot{\forall}p_i(p_i \in \alpha \text{ or } \neg p_i \in \alpha)$ and
- (b) $\dot{\forall}\alpha\dot{\forall}p_i(p_i \notin \alpha \text{ or } \neg p_i \notin \alpha)$.

It is known that $I(A/\Gamma)$ is empty if what is affirmed in A is contained (explicitly or implicitly) in Γ . This takes place if A is a law of C.P.L. since everything that is contained in formulations of laws is contained implicitly in (a) and (b) within the framework of Df 1 and it appears that logical laws do not contain any information (e.g. $I(A \vee \neg A) = \langle M, M \rangle$) and, consequently, the inconsistent proposition $(A \& \neg A)$ contains all information expressible in a given language ($I(A \& \neg A) = \langle \emptyset, M \rangle$). It leads to the following paradoxes of entailment: $(A \& \neg A) \models B$ and $B \models (A \vee \neg A)$.

To obtain a more exact notion of entailment it is necessary to expose the information of A itself without any presuppositions, without any information relative to possible states of affairs in the domain the propositions of the considered language are assigned to (certainly with the exception of that which may be contained in definitions of logical constants; further we shall come across this possibility in semantics of S5 analogous to other modal systems). In other words, it is necessary to exclude presuppositions (a) and (b) from the definition of initial possibilities, i.e., s.d.

Df 3. Bearing in mind an infinite list of propositional variables $p_1, p_2, \dots, p_n, \dots$ in the language, we shall call any subset of the set

$$\{p_1, \neg p_1, p_2, \neg p_2, \dots, p_n, \neg p_n, \dots\}$$

a state description.

Given all other definitions introduced earlier and in particular

$$(2) \quad A \models B \Leftrightarrow M_A \subseteq M_B \Leftrightarrow \dot{\forall}\alpha(TA/\alpha \Rightarrow TB/\alpha),$$

the notion of relevant entailment is obtained $A \models B$. It is natural to introduce also the following generalization:

$$\text{Df 4. } A_1, A_2, \dots, A_m \models B \Leftrightarrow \bigcap_{i=1}^m M_{A_i} \subseteq M_B \quad (m \geq 1).$$

It is clear that $A_1, A_2, \dots, A_m \models B \Leftrightarrow (A_1 \& A_2 \& \dots \& A_m) \models B$ (with any arrangement of parentheses in $A_1 \& A_2 \& \dots \& A_m$). Should we consider the s.d. introduced in Df 3 as descriptions of possible states of affairs or states of our knowledge? The latter is generally believed to be accepted since only theories, opinions, etc., can be inconsistent and incomplete but

states of affairs cannot (ROUTLEY and ROUTLEY, 1972; ANDERSON and BELNAP, 1975, § 16.2.1).

These authors do not feel like referring to inconsistent and incomplete constructions as state descriptions using for them the term "set-ups" introduced by Routley. The above considerations are undoubtedly valid even if to take into account that cognition itself and its results—knowledge, faith, conviction, etc.—side by side with the objective world may be the "reality" itself, that is, "field of knowledge" the propositions p_i , $\neg p_i$ apply to.

Whatever field it may be it is impossible to have some situation p_i in it and at the same time not to have it, $\neg p_i$ (and necessarily, of course, to have one or another). However, the above-mentioned presuppositions that we ought to abstract ourselves from are contained in these very considerations. The fact is that a s.d. set presents abstract possibilities concerned with the abstracting activity of mind. Classical s.d. (at least, when they are used to define logical entailment) are the result of abstracting from specific relations existing in the field of knowledge under consideration and from its laws. This step is associated with an abstraction from the specific content of propositions and with a transition to their logical form consideration only. In relevant logic a further step is necessary: an abstraction both from known to us characteristics of logical nature and from everything that allows us to judge what "can" and what "cannot" be in reality.

The s.d. introduced in DF3 are descriptions of possible states of affairs in reality from the point of view of a human being who knows nothing about reality even of logical character, moreover, that it has or has not situations of different types. Accepting the indicated s.d. we do not reject the presuppositions (a) and (b) at all but abstract ourselves from this knowledge of logical character when we want to define $I(A)$ (similar to the way we abstract ourselves, for example, from what in reality cannot be p and $\neg q$ if " p " stands for "this substance is metal" and " q " stands for "this substance is conductor").

Furthermore, to define a logically valid formula side by side with the indicated general notion of s.d., we retain the notion of *normal s.d.* bearing in mind *classical s.d.*, i.e. satisfying conditions (a) and (b).

DF 5. $\models A$ (the formula A is logically valid) $\Leftrightarrow TA/\alpha$ for any normal s.d. α .

THEOREM. *The set of relations $A \models B$ in the established (relevant) sense is equivalent to the set of formulas $(A \rightarrow B)$ in the theory of first degree entail-*

ment (E_{fde}) system (ANDERSON and BELNAP, 1975, § 15.2) (and in De Morgan's logic).¹

It is natural, however, to introduce also a language analogue $A \rightarrow B$ of the relation $A \models B$.

DF 6. $\models A \rightarrow B$ ($A \rightarrow B$ is valid) $\Leftrightarrow A \models B$ where A and B are formulas of C.P.L. The set of valid formulas of DF5 and DF6 is obviously in agreement with the set of E_{fde} system theorems (first-degree formulas) (ANDERSON and BELNAP, 1975, § 19.2).²

However, other modifications of s.d. notion and, therefore, the relation " \models " are possible on the condition that definitions of logical connectives are preserved (DF1). Thus, accepting the definition of s.d. (DF3) with addition either of the condition $\dot{\forall}\alpha\dot{\forall}p_i(p_i \in \alpha \dot{\vee} \neg p_i \in \alpha)$ (condition (b)) from two of them—(a) and (b)—for classical s.d.) or of the condition $\dot{\forall}\alpha\dot{\forall}p_i(p_i \in \alpha \dot{\vee} \neg p_i \in \alpha)$ (condition (a)), we obtain the set of relations $A \models B$ coinciding with the set of formulas ($A \rightarrow B$) accordingly of Hao Wang logic and dual to it.³ (In the first case the relations $A \models B \vee \neg B$ are obviously excluded for propositions of the type $B \vee \neg B$ become informative, in the second case the relation $A \& \neg A \models B$ is excluded since $A \& \neg A$ contains some limited (but not all possible) information. Adding the condition

$$\dot{\forall}\alpha(\dot{\exists}p_i(p_i \in \alpha \& \neg p_i \in \alpha) \Rightarrow \dot{\forall}p_i(p_i \in \alpha \dot{\vee} \neg p_i \in \alpha))$$

to DF3 the set of relations $A \models B$ coincides with the set of valid formulas $A \rightarrow B$ in Łukasiewicz's three-valued logic⁴ where A and B do not contain " \rightarrow ".

It is certainly clear that modifications of the relation may be conditioned by modifications of the logical constants and introduction of new logical constants as example, in S5.

S5 Semantics

We imply the language of C.P.L. with the addition " \square " and corresponding extension of the formula notion.

DF 7. A state description α is any subset of the set

$$\{\square p_1, \square \neg p_1, p_1, \dots, \square p_n, \square \neg p_n, p_n, \dots\}$$

¹ The formulation of the system is given in Appendix.

² The author cannot prove this assertion.

³ The result is obtained by H. Sanches.

⁴ See Appendix.

complying with the conditions: (1) $\square p_i \in \alpha \Rightarrow p_i \in \alpha$, (2) $\square \neg p_i \in \alpha \Rightarrow p_i \notin \alpha$ for every $i = 1, 2, \dots, n, \dots$ " p_i " stands for "it is necessary that p_i is present", " $\neg p_i$ "—"it is necessary that p_i is absent".

DF 8. An *accessibility relation*: $R\alpha\beta$ (s.d. β is accessible from s.d. α) $\Leftrightarrow \alpha_{\square} = \beta_{\square}$, where α_{\square} (β_{\square}) is a set of elements $\square p_i$ and $\square \neg p_i$ (for every i).

If we define " p_i is contingent in α " (" $\neg p_i$ is contingent in α ") as " $p_i \in \alpha$ and $\square p_i \notin \alpha$ " (" $p_i \notin \alpha$ and $\square \neg p_i \notin \alpha$ "), the accessibility relation apparently means that a contingent situation can appear and disappear.

DF 9. Interpretation. For any α from some set M_r :

$$Tp_i/\alpha \Leftrightarrow p_i \in \alpha; \quad Fp_i/\alpha \Leftrightarrow p_i \notin \alpha.$$

For formulas $(A \& B)$, $(A \vee B)$, $\neg A$ as in C.P.L.

$$T\square A/\alpha \Leftrightarrow \dot{\forall}\beta(R\alpha\beta \Rightarrow TA/\beta); \quad F\square A/\alpha \Leftrightarrow \dot{\exists}\beta(R\alpha\beta \& FA/\beta).$$

DF 10. $\mid_{\overline{S5}} A$ (A is S5-valid) $\Leftrightarrow \dot{\forall}\Gamma\dot{\forall}\alpha(\alpha \in M_r)$.

A relation of consequence is determined by Correlation (1) that is equivalent to

$$A \models B \Leftrightarrow \dot{\forall}\Gamma\dot{\forall}\alpha(\alpha \subset M_r \Rightarrow TA/\alpha \Rightarrow TB/\alpha).$$

THEOREM (on adequacy of the described semantics for system S5). $\mid_{\overline{S5}} A \Leftrightarrow A$ is a theorem of S5; $A \models B \Leftrightarrow \mid_{\overline{S5}} \square(\neg A \vee B)$.

Paradoxicality of the above defined entailment occurs in virtue of pre-suppositions analogous to (a) and (b) for C.P.L. such that any situation $\square p_i$, $\square \neg p_i$, p_i is present or not in α and cannot be both.

To pass to the "relevant" version it is necessary to introduce into s.d. designations for the absence of situations $\neg \square p_i$, $\neg \square \neg p_i$, $\neg p_i$, and formulate explicitly the above-mentioned conditions. In such a case we have the following definition of s.d.

DF 7 (a). *State description* α is any subset of the set $\{\square p_1, \neg \square p_1, \square \neg p_1, \neg \square \neg p_1, p_1, \neg p_1, \dots, \square p_n, \neg \square p_n, \square \neg p_n, \neg \square \neg p_n, p_n, \neg p_n, \dots\}$ satisfying the conditions, for each $i = 1, 2, \dots, n, \dots$:

- (1) $\square p_i \in \alpha \Rightarrow p_i \in \alpha;$
- (2) $\square \neg p_i \in \alpha \Rightarrow \neg p_i \in \alpha;$
- (3) $\neg p_i \in \alpha \Rightarrow \neg \square p_i \in \alpha;$
- (4) $p_i \in \alpha \Rightarrow \neg \square \neg p_i \in \alpha;$ and for each \mathcal{P}_i from $\{\square p_i, \square \neg p_i, p_i\}$;
- (5) $\mathcal{P}_i \in \alpha \dot{\vee} \neg \mathcal{P}_i \in \alpha;$
- (6) $\mathcal{P}_i \notin \alpha \dot{\vee} \neg \mathcal{P}_i \notin \alpha$

(apparently, (3) and (4) are the results of substituting " $p_i \notin \alpha$ ", " $\Box p_i \notin \alpha$ ", " $\Box \neg p_i \notin \alpha$ " by " $\neg p_i \in \alpha$ " " $\neg \Box p_i \in \alpha$ " and " $\neg \Box \neg p_i \in \alpha$ " correspondingly in counterpositions (1) and (2)).

The item related to Fp_i/α in DF9 is replaced by $Fp_i/\alpha \Leftrightarrow \neg p_i \in \alpha$. All other definitions are preserved except that α_{\Box} (and β_{\Box}) in DF8 now is the set of elements $\Box p_i$, $\Box \neg p_i$, $\neg \Box p_i$, $\neg \Box \neg p_i$ of s.d. $\alpha(\beta)$.

The theorem on adequacy of a given semantics in S5 is also valid. But if now in correlation (I) we mean s.d. not restricted by conditions (5) and (6), then the correlation determines relevant relation of entailment for propositions S5 (under these conditions we have usual not relevant \models).

Conditions (1)–(4) may be excluded from the definition of s.d. (DF 7(a)) and derived from the definitions if we generalize the item in DF9 referring to Tp_i/α and Fp_i/α substituting it by $T\mathcal{P}_i/\alpha \Leftrightarrow \mathcal{P}_i \in \alpha$; $F\mathcal{P}_i/\alpha \Leftrightarrow \neg \mathcal{P}_i \in \alpha$ where \mathcal{P}_i is any of the atoms p_i , $\Box p_i$, $\Box \neg p_i$ (for any $i \geq 1$) (with conditions (1)–(4), the equivalences $T\mathcal{P}_i/\alpha \Leftrightarrow \mathcal{P}_i \in \alpha$ and $F\mathcal{P}_i/\alpha \Leftrightarrow \neg \mathcal{P}_i \in \alpha$ are derived as consequences).

The described semantics for S5 is characterized (as for C.P.L.) as informally implying quite definite intuition regarding the sense of the atoms (elements of s.d.) $\Box p_i$, $\Box \neg p_i$ (as the rest, of course). Naturally, modal logic must take the concepts of modality regarding elementary propositions for initial (as the notions of "truth" and "falsity") and to lay claim only to clearing the sense (in informal semantics) or, at least, conditions of truth and falsity of complex propositions. It is evident that in any semantics the sense of propositions is brought out to the same extent as the primary concepts intuitively clear. In any case, however, it is possible to start from the definition of " \Box " in S. Kripke's semantics and to consider the described semantics as a specification of the corresponding "possible worlds" semantics. It may be considered informally at least in that here we have specific "worlds"—s.d. (evidently, very simple) and accessibility relations.

Appendix

I. System E_{fde} (ANDERSON and BELNAP, 1975, § 15.2) (A , B , C are formulas which do not contain " \rightarrow ")

AXIOM SCHEMATA

1. $A \& B \rightarrow A$,
2. $A \& B \rightarrow B$,

3. $A \rightarrow A \vee B$,
4. $B \rightarrow A \vee B$,
5. $A \& (B \vee C) \rightarrow (A \& B) \vee C$,
6. $A \rightarrow \neg \neg A$,
7. $\neg \neg A \rightarrow A$.

RULES

- $R_1.$
$$\frac{A \rightarrow B, B \rightarrow C}{A \rightarrow C}$$
 (from $A \rightarrow B$ and $B \rightarrow C$ to infer $A \rightarrow C$),
- $R_2.$
$$\frac{A \rightarrow B, A \rightarrow C}{A \rightarrow B \& C}$$
,
- $R_3.$
$$\frac{A \rightarrow C, B \rightarrow C}{A \vee B \rightarrow C}$$
- $R_4.$
$$\frac{A \rightarrow B}{\neg B \rightarrow \neg A}$$
.

II. *Hao-Wang's logic* (ERMOLAEVA and MUCHNIK, 1974) (A , B , C are formulas which do not contain " \rightarrow ")

AXIOM SCHEMATA

1. $A \& B \rightarrow A$,
2. $A \& B \rightarrow B \& A$,
3. $A \rightarrow A \vee B$,
4. $A \vee B \rightarrow B \vee A$,
5. $A \& (B \vee C) \rightarrow (A \& B) \vee (A \& C)$,
6. $A \rightarrow \neg \neg A$,
7. $\neg \neg A \rightarrow A$,
8. $\neg(A \& B) \rightarrow \neg A \vee \neg B$,
9. $\neg A \vee \neg B \rightarrow \neg(A \& B)$,
10. $\neg(A \vee B) \rightarrow \neg A \& \neg B$,
11. $\neg A \& \neg B \rightarrow \neg(A \vee B)$,
12. $A \& \neg A \rightarrow B$.

RULES

- $R_1.$
$$\frac{A \rightarrow C, B \rightarrow C}{A \rightarrow C}$$

$$R_2. \frac{A \rightarrow B, A \rightarrow C}{A \rightarrow B \& C}$$

$$R_3. \frac{A \rightarrow C, B \rightarrow C}{A \vee B \rightarrow C}$$

Dual system to Hao-Wang's logic is obtained from II by substitution of Axiom 12 by $A \rightarrow B \vee \neg B$ [6].

III. *First degree fragment of Łukasiewicz's logic* is obtained from II by substitution of Axiom 12 by $A \& \neg A \rightarrow B \vee \neg B$ (ЕРМОЕВА and МУЧЛИК, 1974).

IV. *De Morgan's logic* is obtained from II excluding Axiom 12 (ЕРМОЕВА and МУЧЛИК, 1974).

References

- ANDERSON, A. R., and N. D. BELNAP, Jr., 1975, *Entailment. The logic of relevance and necessity*, vol. I (Princeton University Press, Princeton)
- VOISHVILLO, E. K., 1976, (Войшвилло, Е. К.), *Семантическая информация. Понятия экспенсиональной и экспенсиональной информации*, Сб. „Кибернетика и современное научное познание“ („Наука“, Москва)
- VOISHVILLO, E. K., 1978, (Войшвилло, Е. К.), *Логическое следование, связи и законы логики*, Сб. „Модальные и интенсиональные логики“, тезисы координационного совещания, Москва, июнь 5–7, 1978 (ротапринт) (Москва)
- DUNN, J. M., 1976, *Intuitive semantics for first-degree entailment and “coupled-trees”*, Philosophical Studies, vol. 29, No 3
- ROUTLEY, R., and V. ROUTLEY, 1972, *Semantics of first degree entailment*, Noûs, vol. 6
- ЕРМОЕВА, Н. М., and А. А. МУЧНИК, 1974, (Ермоляева, Н. М., Мучник, А. А.), *Модальные расширения логических исчислений типа Хао Вана*, Сб. „Исследования по формализованным языкам и неклассическим логикам“ („Наука“, Москва)

REFERENTIAL MATRIX SEMANTICS FOR PROPOSITIONAL CALCULI

RYSZARD WÓJCICKI

Institute of Philosophy and Sociology, Polish Academy of Science, Warszawa. Poland

Abstract. The key notion defined and examined in this paper is that of a referential matrix for propositional logical calculi. A necessary and sufficient condition for a propositional logic to possess an adequate referential semantics, i.e. consisting of referential matrices, is stated.

1. Preliminaries

Let \mathcal{L} be a propositional language, and let F_1, \dots, F_n be all *propositional connectives* it involves. We shall denote the set of all formulas of \mathcal{L} by L , and identify \mathcal{L} with the algebra of formulas (L, F_1, \dots, F_n) freely generated by the *propositional letters* p_1, \dots, p_t, \dots

Observe that if $\alpha \in L$ and e is an endomorphism of \mathcal{L} , then $e\alpha$ is the formula that results from α by replacing simultaneously all occurrences of the variables p_1, p_2, \dots in \mathcal{L} by ep_1, ep_2, \dots , respectively. In what follows, endomorphisms of \mathcal{L} will be referred to as *substitution functions* or *substitutions*, for short. A set X of formulas of \mathcal{L} will be said to be *closed under substitutions* iff $eX \subseteq X$, for all substitutions e . By $Sb(X)$ we shall denote the least superset of X closed under substitutions. The sets of formulas of the form $Sb(X)$ will be referred to as *logical systems*.

The term ‘logic’ is sometimes applied as a synonym of the term ‘logical system’. In what follows the denotations of these two terms will be treated as distinct. Loosely speaking, a logic should be viewed as a set of inference rules rather than any set of formulas. In order to provide a more precise definition of that notion let us recall that an operation $C: P(L) \rightarrow P(L)$

$(P(L)$ being the power set of $L)$ is said to be a *consequence operation* on \mathcal{L} (a *consequence* for short) iff

$$(C_1) \quad X \subseteq C(C(X)) \subseteq C(X) \subseteq C(X \cup Y)$$

for all $X, Y \subseteq L$. If, moreover,

$$(C_2) \quad eC(X) \subseteq C(eX)$$

for all substitutions e , the consequence C is called *logical* or *structural*.

By a (*propositional*) *logic* we shall understand a couple (\mathcal{L}, C) where \mathcal{L} is a propositional language and C is a structural consequence defined on \mathcal{L} . For each such consequence C , $C(\emptyset)$ is easily seen to be a logical system. It will be referred to as the *logical system of the logic* (\mathcal{L}, C) . Needless to say that the same logical system may correspond to different logics $(C_1(\emptyset) = C_2(\emptyset)$ does not imply that $C_1 = C_2$).

The class of all structural consequences defined on \mathcal{L} forms a complete lattice with the lattice ordering \leqslant defined by

$$C_1 \leqslant C_2 \quad \text{iff} \quad C_1(X) \subseteq C_2(X)$$

for all $X \subseteq L$. As a matter of fact, the lattice of structural consequences is a complete sublattice of the lattice of all consequences. As it is customary, given any class F of consequence operations defined on \mathcal{L} , we shall denote the least upper bound (supremum) of F by $\sup F$ and the greatest lower bound (infimum) by $\inf F$.

Assume that T is a non-empty set of indices (points of reference, possible worlds) and assume that a certain set H of functions of the form $v: L \times T \rightarrow \{0, 1\}$ is selected to play the role of admissible valuations for \mathcal{L} , i.e. the valuations that conform to the intended meaning of F_i 's. Given any $v \in H$, $v(\alpha, t) = 1$ ($= 0$) reads: *the formula α is true (is false) at the point t under v* .

The set H determines uniquely a structural consequence operation Cn_H on \mathcal{L} defined as follows:

$$(H) \quad \alpha \in Cn_H(X) \text{ iff for each } v \in H \text{ and each } t \in T,$$

$$v(\alpha, t) = 1 \quad \text{whenever} \quad v(\beta, t) = 1 \text{ for all } \beta \in X.$$

If $H = \text{Hom}(\mathcal{L}, \mathcal{A})$, for some algebra \mathcal{A} similar to \mathcal{L} , then I shall write Cn_A instead of Cn_H , and occasionally denote T by T_A . Observe that if $H = \text{Hom}(\mathcal{L}, \mathcal{A})$, the elements of \mathcal{A} are functions of the form $r: T \rightarrow \{0, 1\}$. Let us call such algebras *referential*, more exactly *two-valued referential algebras*.

One may define the consequence Cn_H in still another way. Any couple (\mathcal{A}, D) , where \mathcal{A} is an algebra similar to \mathcal{L} and $D \subseteq P(\mathcal{A})$, is said to be a *generalized logical matrix* for \mathcal{L} (cf. WÓJCICKI, 1973). Given any such matrix (\mathcal{A}, D) , we define the consequence operation $Cn_{(\mathcal{A}, D)}$ determined by this matrix as follows:

- (M) $\alpha \in Cn_{(\mathcal{A}, D)}(X)$ iff, for each $h \in \text{Hom}(\mathcal{L}, \mathcal{A})$ and for each $I \in D$,
 $h\alpha \in I$ whenever $h\beta \in I$ for all $\beta \in X$.

Let \mathcal{A}_W be a referential algebra, T_W being the set of its indices. Define D_W to be the family of all sets of the form

$$D_t = \{r \in \mathcal{A}_W : r(t) = 1\},$$

$t \in T_W$, and put $W = (\mathcal{A}_W, D_W)$. Then, as one may easily verify,

$$Cn_{\mathcal{A}_W} = Cn_W.$$

The matrices of the form $W = (\mathcal{A}_W, D_W)$, where \mathcal{A}_W is a referential algebra and D_W is defined as above, will be called *referential*, more precisely *two-valued referential matrices*.

The important thing about logical matrices is that for each structural consequence operation C on \mathcal{L} there exists a logical matrix M such that $C = Cn_M$ (cf. WÓJCICKI, 1973). Thus generalized logical matrices provide us with a universal tool for examining propositional logic. Clearly, in general, one should not expect that the same universal role can be played by referential matrices. In what follows we shall discuss the matter in more detail.

2. Two-valued referential logics

Let K be a set of referential matrices for \mathcal{L} . Put

$$Cn_K = \inf\{Cn_W : W \in K\}.$$

The consequence Cn_K will be said to be *two-valued referential consequence operation on \mathcal{L} determined by K* , and the couple (\mathcal{L}, Cn_K) will be said to be a *two-valued referential logic*. If $C = Cn_K$, then K will be called a *referential semantics strongly adequate for C* (K is occasionally called *weakly adequate for C* if $C(\emptyset) = Cn_K(\emptyset)$).

The following holds true.

PROPOSITION 1. *A propositional logic (\mathcal{L}, C) is referential iff there exists a single referential matrix W such that $C = Cn_W$.*

PROOF: Obviously, we have to prove only the “if ... then” part of the proposition. Assume that $K = \{W_\sigma : \sigma \in \Sigma\}$ is a referential semantics strongly adequate for (\mathcal{L}, C) . For each matrix W_σ denote the set of indices of the matrix (i.e. of the algebra \mathcal{A}_{W_σ}) by T_σ . Write

$$T = \bigcup \{T_\sigma \times \{\sigma\} : \sigma \in \Sigma\}.$$

The set T will serve as the set of indices of the matrix W we are going to define.

The set A_W of elements of the algebra \mathcal{A}_W will be defined to be a set of functions $r : T \rightarrow \{0, 1\}$ of a certain specific kind. For each function $r : T \rightarrow \{0, 1\}$ and each $\sigma \in \Sigma$ define a function $r/\sigma : T_\sigma \rightarrow \{0, 1\}$ by the condition

$$r/\sigma(t) = r(t, \sigma).$$

(Note that the elements of T are couples of the form (t, σ) , where $t \in T_\sigma$.) We put

$$A_W = \{r : r/\sigma \in \mathcal{A}_{W_\sigma}, \text{ for all } \sigma \in \Sigma\}.$$

In order to complete the construction of the algebra \mathcal{A}_W , and thus the matrix W as well, we have to define operations on A_W corresponding to the connectives of \mathcal{L} . Let $\mathcal{L} = (L, F_1, \dots, F_n)$ and let f_i^σ be the operation corresponding to F_n in \mathcal{A}_{W_σ} . For each selection r_1, \dots, r_k of elements of A_W and each (t, σ) in T put

$$f_i(r_1, \dots, r_k)(t, \sigma) = f_i^\sigma(r_{1/\sigma}, \dots, r_{k/\sigma})(t).$$

The algebra

$$\mathcal{A}_W = (A_W, f_1, \dots, f_k)$$

is a referential algebra, and one may easily verify that $Cn_K = Cn_W$, which concludes the proof.

Observe that the algebra \mathcal{A}_W has been made by “pasting together” the algebras \mathcal{A}_{W_σ} . Since the procedure can be applied to any family of referential algebras of the same similarity type, we have defined a certain operation \oplus that applied to such a family produces a single referential algebra. In view of one-to-one correspondence between referential algebras and referential matrices, \oplus may also be viewed as an operation on matrices. Thus we are allowed to write both

$$\mathcal{A}_W = \bigoplus \{\mathcal{A}_{W_\sigma} : \sigma \in \Sigma\},$$

and

$$W = \bigoplus \{W_\sigma : \sigma \in \Sigma\}.$$

If $\alpha, \varphi \in L$, then the symbol $\varphi(\alpha/p)$ will denote the formula that results from φ by replacing every occurrence of the propositional letter p by α . Clearly, if p does not appear in φ , then $\varphi(\alpha/p) = \varphi$. Note also that $\varphi(\alpha/p) = e\varphi$, where e is a substitution such that $ep = \alpha$.

Let C be a structural consequence on \mathcal{L} . Write $\alpha \sim_c \beta$ whenever $C(\alpha) = C(\beta)$, and write $\alpha \approx_c \beta$ whenever $C(\varphi(\alpha/p)) = C(\varphi(\beta/p))$ for all $\varphi \in \mathcal{L}$ and all propositional letters p . When the relations \sim_c and \approx_c coincide, $\sim_c = \approx_c$, we shall say that the consequence C (or equivalently, the logic (\mathcal{L}, C)) is *self-extensional*.

THEOREM 1. *A propositional logic is referential iff it is self-extensional.*

PROOF: (\rightarrow) Assume that $C = C_{n_W}$ for some referential matrix W . The existence of such a matrix is guaranteed by Proposition 1. Suppose that for some $\alpha, \beta \in L$, $\alpha \sim_c \beta$. It immediately follows that for each valuation h in W and for each $t \in T_W$ (T_W being the set of indices of W), $(ha)(t) = (h\beta)(t)$. Indeed, let for some t , $(ha)(t) = 1$ and $(h\beta)(t) = 0$. Then $\beta \in C(\alpha)$, which contradicts $\alpha \sim_c \beta$.

We have then $ha = h\beta$ for all valuations h which, clearly, yields $C(\varphi(\alpha/p)) = C(\varphi(\beta/p))$ for all φ and p , i.e. $\alpha \approx_c \beta$. Since $\alpha \approx_c \beta$ implies $\alpha \sim_c \beta$, we obtain $\sim_c = \approx_c$.

(\leftarrow) Select Δ to be an arbitrary *basis* for C , i.e. a family S of sets of formulas of L such that, for each $X \subseteq L$, $C(X) = \bigcap \{Y \in S : X \subseteq Y\}$. For instance, one may put $\Delta = \{C(X) : X \subseteq L\}$. The family Δ will serve as the set of indices of the referential algebra \mathcal{A}_W we are going to define.

For each $\alpha \in L$, define α^4 to be a function from Δ into $\{0, 1\}$ such that for each $X \in \Delta$

$$\alpha^4(X) = 1 \quad \text{iff} \quad \alpha \in X. \quad (1)$$

Define the set $A_W = \{\alpha^4 : \alpha \in L\}$ to be the set of elements of \mathcal{A}_W .

Let F_1, \dots, F_n be the connectives of \mathcal{L} . To each F_i we assign the operation f_i on A_W defined as follows:

$$f_i(\alpha_1^4, \dots, \alpha_k^4) = (F_i(\alpha_1, \dots, \alpha_k))^4, \quad (2)$$

k being the arity of F_i and thus f_i as well.

In order for f_i to be well-defined we must have

$$(F_i(\alpha_1, \dots, \alpha_k))^4 = (F_i(\beta_1, \dots, \beta_k))^4 \quad (3)$$

whenever

$$\alpha_1^4 = \beta_1^4, \dots, \alpha_k^4 = \beta_k^4. \quad (4)$$

Let us then verify whether (4) implies (3).

Since Δ is a basis for C , (4) is equivalent to

$$\alpha_1 \sim_C \beta_1, \dots, \alpha_k \sim_C \beta_k, \quad (5)$$

which gives

$$\alpha_1 \approx_C \beta_1, \dots, \alpha_k \approx_C \beta_k, \quad (6)$$

and consequently yields

$$F_1(\alpha_1, \dots, \alpha_k) \sim_C F_i(\beta_1, \dots, \beta_k). \quad (7)$$

Formula (7), again in view of the assumption that Δ is a basis for C , is equivalent to (3).

Thus, we have defined the algebra

$$\mathcal{A}_W = (A_W, f_1, \dots, f_n),$$

and the only thing left to complete the proof is to prove that $C = Cn_W$, W being the referential matrix associated with \mathcal{A}_W .

Select an arbitrary $\alpha_0 \in L$ and an arbitrary $X_0 \subseteq L$, and assume that

$$\alpha_0 \in C(X_0). \quad (8)$$

Let h be a valuation for \mathcal{L} in W , and let

$$hp_i = \alpha_i^A, \quad (9)$$

for all $i = 1, 2, \dots$. Define $e_h \in \text{End}(\mathcal{L})$ to be a substitution such that

$$e_h p_i = \alpha_i. \quad (10)$$

Since C is structural, we have

$$e_h \alpha_0 \in C(e_h X_0). \quad (11)$$

Now, observe that for each $\alpha \in L$,

$$h\alpha = (e_h \alpha)^A. \quad (12)$$

Indeed, by (9) and (10), identity (12) is valid at least whenever α is a propositional letter. This allows us to apply a recursive argument. Let, for some $\alpha_1, \dots, \alpha_k$, $h\alpha_j = (e_h \alpha_j)^A$, $j = 1, \dots, k$. Consider the formula $F_i(\alpha_1, \dots, \alpha_k)$. We have:

$$\begin{aligned} hF_i(\alpha_1, \dots, \alpha_k) &= f_i(h\alpha_1, \dots, h\alpha_k) \\ &= f_i((e_h \alpha_1)^A, \dots, (e_h \alpha_k)^A) \\ &= (F_i(e_h \alpha_1, \dots, e_h \alpha_k))^A = (e_h F_i(\alpha_1, \dots, \alpha_k))^A. \end{aligned}$$

Since both F_i and α_j are arbitrary, we have proved that (9) yields (12). Suppose that for some $X \subseteq A$

$$(h\alpha)(X) = 1 \quad \text{for all } \alpha \in X_0. \quad (13)$$

This gives, by (12) and (1), that $e_h\alpha \in X$, for all $\alpha \in X_0$. In turn, (11) and the assumption that A is a basis for C yield $e_h\alpha_0 \in X$ or equivalently, again by (12) and (1),

$$(h\alpha_0)(X) = 1. \quad (14)$$

We have established that (13) implies (14), and by Proposition 1 we obtain

$$\alpha_0 \in Cn_W(X_0). \quad (15)$$

Now, assume that $\alpha_0 \notin C(X_0)$. A is a basis for C and thus for some $X \in A$, $X_0 \subseteq X$ and $\alpha_0 \notin X$. Define h to be a valuation of L in W such that $hp_i = (p_i)^A$. This, as we have established (cf. the argument which leads from (9) to (12)), yields $h\alpha = \alpha^A$ for all α , and we obtain $(h\alpha)(X) = 1$ for all $\alpha \in X_0$ and $(h\alpha_0)(X) = 0$. Hence $\alpha_0 \notin Cn_W(X_0)$, and the proof is concluded.

Theorem 1 provides us with a useful criterion for deciding whether a given propositional logic is referential or not and consequently whether it has a referential semantics or not. For instance, Positive Hilbert's Logic, Minimal Johansson's Logic, Intuitionistic Logic, not to mention Classical Two-Valued Logic, are easily seen to be self-extensional and thus referential. Logic with Constructible Falsity and Łukasiewicz's Many-Valued Logics may serve as examples of non-referential logics. In the case of modal logics an answer to the question about referentiality of a particular logic of this kind depends on how its consequence operation is defined. More about that in the next section.

One may happen to be interested in whether for a given logical system S there exists a set of referential matrices K such that $S = Cn_K \emptyset$. Call such a system *referential*. An answer to this question is provided by the following corollary of Theorem 1 which I state omitting the proof.

COROLLARY. *A logical system S is referential iff for all α, β, φ , if $\varphi \in S$, then*

- (i) $\alpha, \beta \in S$

implies

- (ii) $\varphi(\alpha/p) \in S$ iff $\varphi(\beta/p) \in S$.

3. Modal logic

Although we are going to deal with modal logics, they will be not of our primary interest by themselves. The main objective of this section is to give an idea of how referential semantics can be applied in dealing with a particular logic and how it is related to semantics of some other important kind.

Let $\mathcal{L}_\square = (L_\square, \rightarrow, \neg, \square)$ be a propositional language of the similarity type $(2, 1, 1)$, i.e. \rightarrow is a binary while \neg and \square are unary connectives. A logic (\mathcal{L}_\square, C) will be said to be *modal* iff

- (i) $C(\emptyset)$ contains all classical tautologies expressed in terms of \rightarrow and \neg , the symbols \rightarrow , \neg being the implication and negation signs, respectively.
- (ii) For each X , $C(X)$ is closed under *Modus Ponens*, i.e. $\beta \in C(X)$ whenever $\alpha, \alpha \rightarrow \beta \in C(X)$.

By a *modal consequence* I shall mean the consequence of a modal logic, and by a *modal system* a set of formulas of \mathcal{L}_\square of the form $C(\emptyset)$, where C is a modal consequence.

Observe that for each modal system M there exists a least modal consequence C such that $M = C(\emptyset)$. The consequence C will be denoted by C_M , and the modal logic $(\mathcal{L}_\square, C_M)$ will be called *associated* with M .

Let us restrict our further discussion to the following well-known modal systems: E, C, K, T, B, S4, S5 usually referred to as modal logics, cf. e.g. K. SEGERBERG (1971). The remarks we are going to make can easier be expressed in terms of referential algebras than in terms of referential matrices, although everything what is said below has quite obvious matrix counterpart.

Call a referential algebra \mathcal{A} similar to \mathcal{L}_\square standard iff the operations $\rightarrow_{\mathcal{A}}$ and $\neg_{\mathcal{A}}$ of \mathcal{A} , i.e. the operations corresponding to \rightarrow and \neg in \mathcal{L}_\square , respectively, satisfy the following two conditions:

$$(r_1 \rightarrow_{\mathcal{A}} r_2)(t) = 1 \quad \text{iff} \quad r_1(t) = 0 \text{ or } r_2(t) = 1,$$

$$(\neg_{\mathcal{A}} r_1)(t) = 1 \quad \text{iff} \quad r_1(t) = 0.$$

Given a standard algebra \mathcal{A} , define $P_{\mathcal{A}}(T_{\mathcal{A}})$ to be the set of all subsets of $T_{\mathcal{A}}$ of the form $\{t: r(t) = 1\}$, $r \in \mathcal{A}$. Furthermore, define the *neighborhood function* $N_{\mathcal{A}}: T_{\mathcal{A}} \rightarrow P(P_{\mathcal{A}}(T_{\mathcal{A}}))$ of \mathcal{A} and the *alternation relation* $R_{\mathcal{A}}$

of \mathcal{A} as follows:

- (N) $\{t: r(t) = 1\} \in N_{\mathcal{A}}(t_0)$ iff $\square_{\mathcal{A}} r(t_0) = 1$;
- (R) $t_1 R_{\mathcal{A}} t_2$ iff $N_{\mathcal{A}}(t_1) \neq \emptyset$ and $t_2 \in \cap N_{\mathcal{A}}(t_1)$.

We shall say that a standard algebra A is:

regular iff for each t in $T_{\mathcal{A}}$ either $N_{\mathcal{A}}(t) = \emptyset$ or $N_{\mathcal{A}}(t)$ is a filter,

normal iff for each t in $T_{\mathcal{A}}$, $N_{\mathcal{A}}(t)$ is a filter,

reflexive, symmetric, transitive iff $R_{\mathcal{A}}$ is reflexive, symmetric, transitive.

The concepts defined are obvious counterparts of some well-known concepts applied in neighborhood semantics and relational semantics for modal logic. It is worth examining then how referential algebras are related to neighborhood and relational frames.

Call a referential algebra \mathcal{A} and a frame F *equivalent* iff the set of all admissible valuations of formulas of \mathcal{L} in F coincides with $\text{Hom}(\mathcal{L}_{\square}, \mathcal{A})$. One may easily show that for each frame F there exists a referential algebra \mathcal{A} such that F and \mathcal{A} are equivalent. At the same time, in order for the equivalence to take place \mathcal{A} must be standard and moreover full, i.e. it must involve as its elements all functions from $T_{\mathcal{A}}$ into $\{0, 1\}$. Observe that standard referential algebras can be viewed as Boolean frames and, in particular, full referential algebras can be viewed as complete atomic Boolean frames. The connection between Boolean frames and neighborhood frames was examined by M. GERSON (1974).

In view of the observation we have made, it is somewhat surprising that the word 'full' does not appear in the following theorem (I shall omit the proof of it).

THEOREM 2. $C_E, C_C, C_K, C_T, C_B, C_{S4}, C_{S5}$ are respectively determined by the class of all

- (E) standard referential algebras,
- (C) regular referential algebras,
- (K) normal referential algebras,
- (T) normal and reflexive referential algebras,
- (B) normal, reflexive, and symmetric referential algebras,
- (S4) normal, reflexive, and transitive referential algebras,
- (S5) normal, reflexive, symmetric, and transitive referential algebras.

COROLLARY. All logics of the form $(\mathcal{L}_{\square}, C)$ where C is one of the consequences considered in Theorem 2 is referential. Furthermore, $(\mathcal{L}_{\square}, C_E)$ is the weakest referential logic in the sense that C_E is the greatest lower bound of all modal and referential consequences on \mathcal{L}_{\square} .

It may be of some interest to examine how the consequence operations determined by the classes of full referential algebras of the same kind as those defined by clauses (E)–(S5) of Theorem 2 can be characterized in a syntactical way.

4. Many-valued referential matrices

Let us call the algebras whose elements are functions of the form $r: T \rightarrow V$, V being a set of cardinality μ , μ -valued referential algebras. Assume that \mathcal{A}_W is such an algebra. Select a subset $V_0 \subseteq V$ and define D_W to be the family of all sets of the form

$$D_t = \{r \in A_W : r(t) \in V_0\}.$$

The matrix $W = (\mathcal{A}_W, D_W)$ will be called a μ -valued referential matrix (many-valued, occasionally), and V_0 will be called the set of designated elements of it.

The notion just defined is an obvious generalization of that of a two-valued referential matrix and perhaps is worthwhile being examined. Any-way, examples of logics which do not have two-valued referential semantics, but have many-valued referential semantics are known. For instance, three-valued referential semantics with truth, falsity, and “gap” serving as the truth-values and truth being selected as the designated value is strongly adequate (cf. R. THOMASON, 1969, for Logic with Constructive Falsity). As another example may serve (cf. DUNN, 1976) Dunn’s three-valued semantics (with the values $\{F\}$, $\{T\}$ and $\{T, F\}$, the last two being designated) strongly adequate for one of the relevance logics. The logic to which Dunn’s semantics applies is determined by the system RM and the following two inference rules: from A , $A \rightarrow B$ infer B (*modus ponens*), and from A , B infer $A \wedge B$.

References

- DUNN, J. M., 1976, *A Kripke-style semantics for R-single using a binary accessibility relation*, The Journal of Symbolic Logic, vol. 35
- GERSON, M., 1974, *A comparative study of modal propositional semantics* (Simon Fraser University, Burnaby)
- SEGERBERG, K., 1971, *An essay in classical modal logic*, Filosofiska Studier Uppsala University
- THOMASON, R. H., 1969, *A semantical study of constructible falsity*, Zeitschrift für mathematische Logik und Grundlagen der Mathematik, vol. 15
- WÓJCICKI, R., 1973, *Matrix approach in methodology of sentential calculi*, Studia Logica, vol. 32, pp. 7–37

REALISM IN THE NATURAL SCIENCES

ROY BHASKAR

University of Edinburgh, Great Britain

1. Tensions in recent philosophy

Recent philosophy of science wears an air of paradox. The fundamental assumptions of the positivist world view, viz. that science is *monistic* in its development and *deductive* in its structure, lie shattered. But the ensuing accounts of science have not found it easy to sustain a coherent notion of the rationality, or even intelligibility, of either scientific change or the non-deductive component of theory. I think that one can trace the source of this difficulty back to the continuance, alongside the new philosophy of science, of an old philosophy of being, materially incompatible with it. The result is that philosophy is caught in a cleft stick. With the new epistemology it cannot go back. But without a new ontology it cannot go forward. The effects of this tension are clearly visible along both the anti-monistic and anti-deductivist limbs of the anti-positivist pincer.

Consider first the anti-monistic movement, represented most notably perhaps by the work of Bachelard, Koyré, Popper, Lakatos, Feyerabend and Kuhn. Both Bachelard and Kuhn come very close to the position, whose roots lie in Vico, and which I shall characterise as *super-idealism*, that we create and change the world along with our theories (see e.g. BACHELARD, 1936, pp. 63–64, and KUHN, 1970a, p. 121). Neither Kuhn nor Feyerabend have managed to sustain the intelligibility of the concept of a *clash* between incommensurable descriptions, or to say *over* what such descriptions clash. Popper has not shown how the falsification of a conjecture could be rational, *unless* nature were uniform. And he has not furnished any ground for assuming that it is, in the face of Humean and Goodmanesque possibilities.

Nor has Lakatos shown how unless nature were uniform, it would be rational to work on progressive rather than degenerating programmes; or, for that matter, pay any attention to the history of science. More generally, the theorists of scientific change have found it difficult to reconcile the phenomenon of discontinuity with the seemingly progressive, cumulative character of scientific development, in which there is growth as well as change.

Parallel problems beset the anti-deductivist movement. Under the initial influence of Wittgenstein, philosophers such as Hanson, Toulmin, Hesse and Harré have sought to show how scientific practice generates cognitive items—be they glossed as paradigms, heuristics, conceptual schemata, models or ideals—irreducible to syntactical operations upon sense-experience, and which are essential for both the intelligibility and the empirical extension of theory. Such items function, as it were, as surrogates for natural necessity (see HARRÉ, 1973, pp. 358–380). The problem is this: if the surrogate can be empirically described, then its postulation is legitimate, but it now ceases to play any independent role, so that the necessity of the connection, the analogical character of the model, the ideality of the order, etc., vanishes. Conversely, if it cannot be empirically described, its cognitive function is retained, but it now (on the ontology of empirical realism) ceases to explicate the nature of any real phenomenon (cf. e.g. HEMPEL, 1963, Chap. 8). More generally, writers within this tradition have not always succeeded in counterbalancing their stress on the synthesising activity of the scientific imagination with the messy practicalities of science's causal interaction with nature (the nuts and bolts, so to speak, of scientific life).

Now I think that if the rational insights of both the anti-monistic and anti-deductivist tendencies are to be saved, a new ontology must be constructed for them. Such an ontology involves a Copernican Revolution in the strict sense of an anti-anthropocentric shift in our philosophical conception of the place of man in nature. It is my aim in this paper to show the necessity for the new realist philosophy of natural science such a shift entails.

2. Types of realism

Realism is the theory that the ultimate objects of scientific inquiry exist and act (for the most part) quite independently of scientists and their activity. Now, as so defined, it might be thought that the question of whether

or not natural science is ‘realist’ can only be answered empirically, viz. by determining whether or not scientists believe, or act as if they believe, that the theoretical entities and processes they posit possess real referents independent of their theorising (cf. PUTNAM, 1978, pp. 133–140). Such questions are clearly legitimate and necessary. But I want to argue the case for a metaphysical realism, consisting in an elaboration of what the world must be like prior to any scientific investigation of it and for any scientific attitudes or behaviour to be possible. Such a realism neither presupposes nor licenses a realistic interpretation of any particular theory.

Clearly, the possibility of such a metaphysical, as distinct from ‘internal’, realism will depend upon the establishment of the possibility of a *philosophy*, as distinct from sociology (or history) of science. But within philosophy, it will also depend upon the possibility of an *ontology*, as distinct from epistemology. For realism is not a theory of knowledge or of truth, but of *being* (though as such it has of course epistemological implications). Accordingly, a realist position in the philosophy of science will be a theory about the nature of the being, not the knowledge, of the objects investigated by science—roughly to the effect that they exist and act independently of human activity, and hence of both sense-experience and thought. In this way realism is immediately opposed to both empiricism and rationalism; and to that opinion of post-Humean philosophy—which I shall call the *epistemic fallacy*—that ontological questions can always be rephrased in epistemological form: that is, that statements about being can always be analysed in terms of statements about our knowledge (of being), that it is sufficient for philosophy to “treat only the network, and not what the network describes” (cf. Wittgenstein, 1961, 6.35).

Now it is clear that any theory of the knowledge of objects entails some theory of the objects of knowledge; that every theory of scientific knowledge must logically presuppose a theory of what the world is like for knowledge, under the descriptions given it by the theory, to be possible. Thus, suppose a philosopher analyses scientific laws as, or as dependent upon, constant conjunctions of events, he is then committed to the view that there *are* such conjunctions; that, in Mill’s words, “there are such things in nature as parallel cases; that what happens once will, under a sufficient degree of similarity of circumstance, happen again.” (MILL, 1961, Bk. III, Chap. 3, Sect. 1.) In this way, then, as Bachelard recognised, “all philosophy, explicitly or tacitly, honestly or surreptitiously ... deposits, projects or presupposes a reality.” (BACHELARD, 1953, p. 411.) So we could say, inverting a famous dictum of Hegel’s—every philosophy (at least in as much as it is

a philosophy of science)¹ is essentially a realism, or at least has realism for its principle, the only questions being then how far, and *in what form*, this principle is actually carried out (HEGEL, 1965, pp. 154–155). Now the orthodox tradition in the philosophy of science, including both its Humean and Kantian wings, has depended upon an implicit ontology of *empirical realism*, on which the real objects of scientific investigation are the objects of actual or possible experience. More recently, the super-idealist tendency has secreted an implicit ontology of *subjective conceptual realism*, on which the real objects of scientific investigation are the products of scientific theory (that is, of the spontaneous activity of mind, unconstrained by sense-experience). But I want to show that only a realism fully consistent with the principle (or definition) of realism enunciated above, *transcendental realism*, can sustain the intelligibility of the experimental and theoretical work of science.

3. On method

How then is a philosophy of science possible? What distinguishes philosophy from science is not its concern with a special field (e.g. language-culture or man), nor the generality of the questions it asks (whether this is conceived as a matter of degree, as in Quine, or kind, as in Lakatos), nor its investigation of (participation in or contribution to) some autonomous order of being. Rather philosophy distinguishes itself from science by its *method*, and more generally by the kinds of considerations and arguments it deploys, which are transcendental in Kant's sense.

Now although if philosophy is to be possible, it must pursue a transcendental procedure, it must reject the idealist and individualist mould into which Kant pressed his own inquiries. In fact, if the general form of a philosophical investigation is into the necessary conditions of conceptualised

¹ That is, in as much as the philosophy is to be at all *relevant* to the practice of science. As both Hume and Hegel realised, scepticism—in the sense of suspension of commitment to some idea of an independent reality—is not a tenable (or ‘serious’) position. Thus: “whether your scepticism be as absolute and sincere as you pretend, we shall learn by and by, when the company breaks up; we shall then see whether you go out at the door or the window, and whether you doubt if your body has gravity or can be injured by its fall, according to popular opinion derived from our fallacious senses and more fallacious experience”, HUME, 1948, p. 7. And: “(Scepticism) pronounces absolute disappearance and the pronouncement exists...; it pronounces the nullity of seeing, hearing, etc., and it itself sees and hears etc.; it pronounces the nullity of ethical realities, and acts according to them”, HEGEL, 1949, p. 250. Cf. also ENGELS & MARX, 1970, p. 48.

activities, then it must be recognised that both social activity and philosophical conceptualisation may be historically transient; that the activity may depend upon the powers of people as material objects or causal agents rather than merely thinkers or perceivers; and that its analysis may yield transcendental realist, not idealist, and epistemically relativist, rather than absolutist (or irrationalist), conclusions. On this conception, then, both the premisses and conclusions of philosophical arguments remain contingent facts, the former but not the latter being necessarily social (and so historical). It is only in this relative or conditional sense that philosophy can establish synthetic *a priori* truths. For philosophy gets going always (and only) on the basis of prior conceptualisations of historical practice, i.e. of specific ideas of determinate social forms.

Philosophy, then, does not consider a world apart from that of the various sciences. Rather it considers just that world, but from the perspective of what can be established about it by *a priori* argument, where it takes as its premisses scientific activities as conceptualised in experience (or in a theoretical redescription of it). As such, philosophy is *dependent* upon the form of scientific practices, but *irreducible* to the content of scientific beliefs. Thus philosophy can tell us that, if experimental activity is to be intelligible, the world must be structured and differentiated. But it cannot tell us what structures the world contains or the ways in which they are different, which are entirely matters for substantive scientific investigation. If philosophy does not compete with science, in virtue of its transcendental nature, it does not exist apart from science, in virtue of its syncategorematic character. For the terms of a philosophical discourse denote only on the condition that they are used under particular descriptions in science. Thus whatever is philosophically demonstrable is also in principle scientifically comprehensible. And hence in the long run relatively autonomous philosophy must be *consistent* with the findings of science.

But how are we to select the premisses of our transcendental arguments without already implying an unvalidated commitment to the epistemic significance of the activities described? Recourse to an arbitrary and external criterion of knowledge (cf. HEGEL, 1949, Introduction, pp. 131–145) can be avoided by focussing on those activities which non-realists have historically picked out as most significant in science. Thus considering experimentation, sponsored by empiricists and Kantians, and conceptual transformations, sponsored by super-idealists, I will show (i) how the sponsoring theory cannot sustain the intelligibility of the sponsored activity without metaphysical absurdity, and (ii) how a realist analysis can render

the sponsored activity intelligible. I do not claim that my analyses are certain or unique (though they are the only plausible analyses I know of). But they are demonstrably superior to the non-realist alternatives that currently hold the floor in contemporary philosophy. Moreover, the resulting realist account of science provides a clear and consistent alternative to positivism which allows us both to save the cumulative character of science without restoring a monism and to rescue a 'surplus' component in scientific theory without plunging into subjectivism.

4. Experimental activity and the vindication of ontology

For the empiricist experimental activity is necessary, and perhaps sufficient, for the establishment of causal laws and other items of general knowledge; and causal laws etc. are analysed as, or as dependent upon, constant conjunctions of events (or states of affairs) perceived or perceptions. It is not difficult to see that this analysis is faulty.

In an experiment scientists co-determine, i.e. are causally co-responsible for, a pattern of events. There is nothing in itself special about this. For, as causal agents, we are continually co-responsible for events. What is significant about the patterns scientists deliberately produce under conditions which they meticulously control is that it enables them to identify the mode of operation of structures, mechanisms or processes which they do not produce. What distinguishes the phenomena the scientist *actually* produces out of the totality of the phenomena he *could* produce is that, when his experiment is successful, it is an index of what he does *not* produce. A *real* distinction between the objects of experimental investigation, such as causal laws, and patterns of events is thus a condition of the intelligibility of experimental activity. Now as constant conjunctions must in general be artificially produced, if we identify causal laws with them, we are logically committed to the absurdities that scientists, in their experimental activity, cause and even change the laws of nature! Thus the objects of scientific inquiry in an experiment cannot be events and their conjunctions, but are (I suggest) structures, generative mechanisms and the like (forming the real basis of causal laws), which are normally out of phase with them. And it can now be seen that the Humean account depends upon a misidentification of causal laws with their empirical grounds (see BHASKAR, 1975).

But, of course, we not only experimentally establish, we practically *apply* our knowledge—in systems, which may be characterised as *open*, where no constant conjunctions obtain. If this activity is to be rendered intelligible,

causal laws must be analysed as tendencies, which may be possessed unexercised and exercised unrealised, just as they may of course be realised unperceived (or undetected) by men. Thus in citing a law we are referring to the trans factual activity of mechanisms, that is, to their activity as such, not making a claim about the actual outcome (which will in general be co-determined by the effects of other mechanisms too). And a constant conjunction, or empirical invariance, is no more a necessary, than it is a sufficient condition for the operation of a causal law. Here again, failure to mark the ontological difference between causal laws and patterns of events issues in absurdity. For if causal laws are, or depend upon, constant conjunctions, then we must ask: what governs phenomena in open systems, that is in the vast majority of cases? The empiricist is now impaled on an acute dilemma—for he must either aver that nothing does, so that nature becomes radically indeterministic; or suppose that, as yet, science has discovered no laws! (BHASKAR, 1975, Chap. 2.)

Once made, however, the ontological distinction between causal laws and patterns of events allows us to sustain the universality of the former in the face of the non-invariance of the latter. Moreover the Humean analysis of laws now loses all plausibility. For the non-invariance of conjunctions is a condition of an empirical science and the non-empirical nature of laws a condition of an applied (or pragmatic) one.

Did we not know this all along? Of course, it is in line with our intuitions. Thus we do not suppose that e.g. Ohm's Law or Prout's Hypothesis holds only in the laboratory—where alone they can be tested. And as every research worker knows: no experiment goes properly the first time. We can use our knowledge for the explanation of events and the production of things in open systems, where deductively-justified predictions, and decisive test situations, are impossible. And yet in the reflective consciousness of philosophy, as distinct from the spontaneous practice of science, it has seldom been doubted that the Humean analysis specifies at least necessary conditions for the attribution of laws.

Of course, transcendental idealists and others have long contended that a constant conjunction of events is not a sufficient condition for a causal law. They have seen that no scientist ever fails for a moment to distinguish a necessary from an accidental sequence (even if he is not always sure into which class a given sequence falls). But the problem has always been to ground this intuition in such a way as to sustain a concept of *natural* necessity, that is a necessity in nature quite independent of men and their activity. More recently, Anscombe, von Wright and some others, have

noted that our active *interference* in nature is normally a condition of empirical regularities. But they have not seen that it follows from this that there must be an *ontological* distinction between such regularities and the laws they ground. (We produce not the laws of nature, but their empirical grounds.) On the transcendental realist system, a sequence *A*, *B* is necessary if and only if there is a natural mechanism *M* such that when stimulated by *A*, *B* tends to be produced. It is a condition of the experimental establishment and practical application of our knowledge that such mechanisms exist and act, as what may be termed the *intransitive* objects of scientific inquiry, independently of their identification by men. And it is in their transfactual activity—described in ‘normic’ statements—that the real ground for the ‘surplus-element’ in the analysis of laws lies.

The analysis of experimental activity shows that causal laws are ontologically distinct from patterns of events. But experimental activity involves sense-perception (as well as causal agency); and reflection on the necessity for a scientific training (or the possibility of scientific change) shows that events must be ontologically distinct from experiences. The concept of causal laws as, or as dependent upon, empirical regularities thus involves a double reduction: of causal laws to constant conjunctions of events and of such events to experiences. This double reduction involves two category mistakes, expressed most starkly in the concepts of the empirical world and of the actuality of causal laws (which presupposes the ubiquity and spontaneity of closed systems).

Now in a world without men there would be no experiences and few, if any, constant conjunctions of events. For both experiences and invariances depend, in general, upon human activity. But causal laws do not. Thus in a world without men, the causal laws that science has now as a matter of fact discovered would continue to prevail, though there would be few sequences of events and no experiences with which they were in correspondence. The analysis of experimental activity shows, then, that the assertion of a causal law entails the possibility of a *non-human world*, i.e. that it would operate even if it were unknown, just as it continues to operate when its consequent is unrealised (or if it is unperceived or undetected by men), i.e. outside the conditions that permit its empirical identification. It follows from this that statements about being cannot be reduced to or analysed in terms of statements about knowledge, that ontological questions cannot always be transposed into epistemological terms. Thus the transcendental analysis of experience, the empiricist’s criterion of knowledge, establishes both that a philosophical ontology is possible and some propositions in it

(e.g. causal laws are distinct from patterns of events, and events from experiences). But the epistemic fallacy in philosophy covers or disguises an *implicit ontology* based on the category of experience, and an *implicit realism* based on the presumed characteristics of the objects of experience, viz. atomistic events, and their relations, constant conjunctions. From Hume onwards philosophers have thus allowed, for the sake of avoiding ontology, a particular concept of our knowledge of reality, which they may wish to explicitly reject, to inform and implicitly define their concept of the reality known by science. The result has been a continuing '*ontological tension*' induced by the conflict between the rational intuitions of philosophers about science and the constraints imposed upon their articulation by their inherited ontology. This has led to a nexus of interminably insoluble problems (such as the problem of induction), the anthropocentric displacement of these intuitions and the opening up of a fissure between the methodological implications of epistemology and the realist practice of science.

Now if the objects of our knowledge exist and act independently of the knowledge of which they are the objects, it is equally the case that such knowledge as we actually possess always consists in historically specific social forms. Thus to think our way clearly in the philosophy of science we need to constitute a *transitive* dimension or epistemology to complement the intransitive dimension or ontology already established. It is evident that, unless we do so, any attempt to establish the irreducibility of knowable being—which is the only kind of being with which science is concerned—to thought must end in failure.

5. On the epistemology of scientific change

Once an intransitive dimension is established, both new and changing knowledge of independently existing and acting objects becomes possible. Now if we are to avoid the absurdity of the assumption of the production of such knowledge *ex nihilo* (on which more anon), it must depend upon the employment of antecedently existing cognitive materials, which I have called the *transitive* objects, and which function as the material causes, of knowledge. So science must be seen as a social process, irreducible to an individual acquisition, whose aim is the production of the knowledge of the mechanisms of the production of phenomena in nature, the intransitive objects of inquiry.

Now as it is clear that the hypothetical entities and mechanisms imagined for the purposes of theory-construction must initially derive at least part of their meaning from some other source (if they are to be capable of functioning as possible explanations at all) theories must already be understood before correspondence rules are laid down for them. Equally this means that the descriptive terms must initially have possessed a meaning independent of them; so that meaning-change is not only possible, but inevitable in the process of science. Now it clearly could come to pass over some scientific transformation that, as e.g. Feyerabend and Kuhn have suggested, no meanings are shared in common between two conflicting scientific theories. Can we then still sustain the notion of a rational choice between such incommensurable theories? Yes. For we can allow quite simply that a theory T_A is preferable to a theory T_B , even if they are incommensurable, provided that T_A can explain *under its descriptions* almost all the phenomena that T_B can explain under its descriptions *plus* some significant phenomena that T_B cannot explain. Now patently the possibility of saying this depends upon the explicit recognition of a philosophical ontology or intransitive dimension, and this is of course just what the super-idealists deny. But such an ontology is already implicit in the very formulation of the problem, or definition of the phenomenon, of incommensurability. For to say of two theories that they conflict, clash or are in competition presupposes that there is something—a domain or real objects or relations existing and acting independently of their (conflicting) descriptions—*over* which they clash. Hence incommensurable theories must share a part world in common. If they do not, i.e. if the phenomenon of Kuhn-loss is total, then no sense can be given to the concept of scientific change, and *a fortiori* to the notion of a clash between the theories (for they are now no longer alternatives). Such a total replacement involves neither transformation nor discursive intelligence, but an archetypal intuitive understanding constructing its world in a single synthetic act (cf. KANT, 1972, pp. 249–258); and the inexplicable solipsism it entails is devoid of significance for us.

A rational account of scientific development follows on quickly from the establishment of the transcendental realist ontology of structures and differences. Typically the construction of an explanation for, that is the production of the knowledge of the mechanisms of the production of, some identified phenomenon will involve the building of a model, utilising antecedently existing cognitive resources (not already employed in the description of the domain in question) and operating under the control

of something like a logic of analogy and metaphor, (see e.g. HARRÉ, 1970, Chap. 2, and HESSE, 1974, esp. Chaps 9 and 11), of a mechanism, which *if* it were exist to and act in the postulated way would account for the phenomenon in question (a movement of thought which, following Hanson, may be called 'retroduction' (see HANSON, 1965, pp. 85 ff)). The reality of the postulated mechanism must then, of course, be subjected to empirical scrutiny.² (For in general more than one explanation will be consistent with the phenomenon concerned.) Once this is done, the explanation must then in principle itself be explained. And so we have a three-phase schema of development in which, in a continuing dialectic, science identifies a phenomenon (or range of phenomena), constructs explanations for it and empirically tests its explanations, leading to the identification of the generative mechanism at work, which now becomes the phenomenon to be explained; and so on. If the classical empiricist tradition restricts itself to the first phase, the neo-Kantian tradition sees the need for the second, but it either denies the need for, or does not draw the full implications of, the third. Transcendental realism differentiates itself from empirical realism in interpreting the first phase of the dialectic as the invariance of a *result* rather than a *regularity* and from transcendental idealism in allowing that what is *imagined* at the second need not be *imaginary* but may be (and come to be known as) *real*. Now in this continuing process, as deeper levels or strata of reality are successively unfolded, scientists must construct and test their explanations with the cognitive resources and physical tools at their disposal, which in this process are themselves progressively transformed, modified and refined.

On the transcendental realist view of science, then, its essence lies in the *movement* at any one level from knowledge of manifest phenomena to knowledge, produced by means of antecedent knowledge, of the structures that generate them. Now knowledge of deeper levels may correct, as well as explain, knowledge of more superficial ones. In fact one finds a characteristic pattern of description, explanation and redescription of the phenomena identified at any one level of reality. But only a concept of ontological depth (depending upon the concept of real strata apart from our knowledge of

² It is important to note that science employs two criteria for the ascription of reality to a posited object: a perceptual criterion and a causal criterion. The causal one turns on the capacity of the entity to bring about changes in material things. Notice that a magnetic or gravitational field satisfies this criterion, but not a criterion of perceptibility. On this criterion, to be is not to be perceived, but rather (in the last instance) just to be able to do.

strata) enables us to reconcile the twin aspects of scientific development: growth and change. And hence both to avoid the onesidedness of the accounts of continuists, such as Nagel, and discontinuists, such as Popper, alike; and to sustain (in opposition to e.g. Feyerabend and Kuhn) the rationality of scientific transformations. Moreover, only the concept of ontological depth can reveal the actual historical stratification of the sciences as anything other than an accident. For this can now be seen as grounded in the multi-tiered stratification of reality, and the consequent logic—of discovery—that stratification imposes on science.

This logic must be located in the movement or transition from the identification of invariances to the classification of the structures or mechanisms that account for them. In this transition, Humean, Lockean and Leibnizian knowledge of the objective world-order is progressively obtained. At the first (Humean) level, we just have the invariance of an experimentally produced result. Given such an invariance, science moves immediately to the construction and testing of possible explanations for it. If there is a correct explanation, located in the nature of the thing or the structure of its system, then there is a reason independent of its actual behaviour for that behaviour. Such a reason may be discovered empirically. And, if we can deduce the things normic behaviour from it, then the most stringent possible (or Lockean) criterion for our knowledge of natural necessity is satisfied. For example, we may discover that copper has a certain atomic or electronic structure and then be able to deduce its dispositional properties from a statement of that structure. If we can do so, we may then be said to possess knowledge of natural necessity *a posteriori*. Finally, at the third (or Leibnizian) level, we may seek to express our discovery of its structure in an attempted real definition of the substance, process or thing. (Causal laws then appear as the tendencies of natural kinds, realised under closed conditions.) This is not to put an end to inquiry, but a stepping stone to a new process of discovery in which science seeks to unearth the mechanisms responsible for *that* level of reality.

It is clear that for an adequate account of scientific development both the concepts of a stratified and differentiated reality and of knowledge as a produced means of production must be sustained. A critique of empiricism is achieved by noting how knowledge at the Lockean level, viz. of real essences, is possible, so resolving the paradoxes and problems (most notoriously, of induction) that stem from the dogmatic postulation or unthinking assumption of empirical realism. But a complementary critique

of rationalism is achieved by noting that such knowledge is produced, in the context of a dialectic of explanatory and taxonomic knowledge, *a posteriori*—in the transitive, irreducibly empirical process of science.

6. Philosophies as ideologies of science

Now the orthodox tradition in the philosophy of science, including both its empiricist and neo-Kantian wings, has uncritically accepted the doctrine, implicit in the empirical realist dissolution of ontology, of the actuality of causal laws; and it has interpreted these, following Hume, as empirical regularities. In this way, by secreting an ontology based on the category of experience, three domains of reality (the domains of the real, the actual and the empirical) are collapsed to one. Now this double reduction prevents the empirical realist from examining the critical question of the conditions under which experience is *in fact* significant in science. In general this depends upon the transformation of both man and nature, so that the per-
cipient is skilled and the system in which the phenomenon occurs is closed. It is only when the distinctiveness of the domains is registered, and the possibility of their disjuncture thereby posed, that we can appreciate the enormous effort—in experimental design and scientific education—required to make human experience epistemically significant in science. (Research and teaching are the two most obvious, yet philosophically underanalysed, tasks of scientists, just as the laboratory and the classroom are the two most obvious *sites* of science.)

It is evident that the critical omission from orthodox accounts of science is the notion of scientific activity as *work*. Moreover when, as in transcendental idealism, work is recognised, it is treated only as intellectual, and not also as practical labour, in causal exchange with nature. Accordingly, such accounts cannot see knowledge, or at least the achievement of a closure, as a transient social product. Underlying the undifferentiated ontology of empirical realism is thus an individualistic sociology, in which people are regarded as passively sensing (or else, as conventionally deciding upon) given facts and recording their constant conjunctions, i.e. as passive spectators of a given world, rather than as active agents in a complex one. In the ensemble of conditions and concerns that constitute empirical realism, it is this model of man that plays the dominant role. For it is the need felt by the philosophy of science, conceiving its role as the guarantor of justified belief (rather than as the analyst of intelligible activities), for certain foun-

dations for scientific knowledge that determines the atomicity of experiences, and hence of their ontological counterparts, which in turn necessitates the constancy of their conjunctions, i.e. the closure of the systems within which the events occur.

It can thus be seen that the complement of the anthropocentricity implicit in the empiricist analysis of laws, and necessary for it, is neglect of the conscious human activity required for our knowledge of them. For both experiences, together with the facts they ground, and the conjunctions that, when apprehended in sense-experience, provide the empirical grounds for laws, are social products. But the Humean theory depends upon a view of conjunctions existing quite independently of the human activity necessary for them, and hence upon the *fetishism* of the systems within which the conjoined events occur. And it depends upon a view of what is apprehended in immediate sense-experience as a fact constituting an atomistic event or state-of-affairs, and existing independently of the human activity necessary for it, and hence upon the *reification* of atomised facts, apprehended by autonomised minds. When the conjunctions of such facts are reified and identified with causal laws, science becomes an epiphenomenon of nature. Thus, in the intellectual grid within which philosophical ideas are produced, the man-dependence of knowlege (its social nature) and the man-independence of the world (its transcendentally real character), appear in empirical realism as the man-dependence of the world (its empirical nature) and the activity-independence of knowledge (its a-social character). In this way, a naturalised science is purchased at the price of a humanised nature; and the concept of the empirical world finds its counterpart and condition in a reified account of science.

The effects of these transformations are striking. The positivistic concept of a fact as what is more or less immediately apprehended in sense-perception generates characteristic ideologies *for* and *of* science. The former rationalises the practice of what Kuhn has called 'normal science'; while the latter secretes mystiques of commonsense and/or expertise. Similarly, descriptivist, instrumentalist and fictionalist interpretations of theory, by reducing the ontological import of theories to a given self-certifying experience, serve to exempt our current claims to theoretical knowledge from criticism. Or again, to consider a more general effect, the Humean theory of causality, presupposing a view of the world as closed and completely described, encourages a conception of the social world as unstructured (hence as 'obvious'), undifferentiated and unchanging, so underpinning certain substantive theories of social life.

If empirical realism involves reification and rationalises normal science, the super-idealist ontology of subjective conceptual realism involves a *voluntarism*, on which theory is unconstrained by either nature or history, which readily lends itself to the rationalisation of so-called 'revolutionary science'. Of course, both ideologies possess a measure of partial adequacy—in that they accord with our *spontaneous* consciousness in science. Thus we do tend to read the world *as if* it were constituted by facts, rather than particulars, in 'epistemic perception' (cf. DRETSKE, 1969, Chap. 1); and in moments of creativity, we experience ideas as coming 'out of the blue' or, as we say (in defiance of the First Analogy), from nowhere.

7. Some implications of realism

In conclusion, I want to indicate briefly some of the implications of the new realist ontology and account of science.

Transcendental realism explicitly asserts the non-identity of the objects of the transitive and intransitive dimensions, i.e. of thought and being. And it relegates the notion of a correspondence between them to the status of a metaphor for the aim of an *adequating practice* (in which cognitive matter is worked into a matching representation of a non-cognitive object). It entails acceptance of (i) the principle of *epistemic relativity*, viz. that all beliefs are socially produced, so that all knowledge is transient, and neither truth-values nor criteria of rationality exist outside historical time. But rejection of (ii) the doctrine of *judgmental relativism*, which maintains that all beliefs are equally valid, in the sense that there can be no rational grounds for preferring one to another. It thus stands opposed to epistemic absolutism and epistemic irrationalism alike. Relativists have wrongly inferred (ii) from (i) (see e.g. KUHN, 1970a, pp. 264–265), while anti-relativists have wrongly taken the unacceptability of (ii) as a reduction of (i) (see e.g. POPPER, 1972, p. 308).

By making the possibility of philosophical discourse contingent upon the actuality of social practices, transcendental realism provides a way of integrating philosophical and sociological (or historical) studies of practices such as science. Moreover, through the resolution of the problems generated by the notion of the contingency of the causal connection and the critique of the deductivist (and deterministic) theories generated by the notion of its actuality, the scene is set for a philosophy that will once more act as 'underlabourer' (cf. LOCKE, 1959, p. 14), and occasional midwife, to the

sciences. On the new world view that emerges both nature and the sciences are stratified and differentiated; and the possibility arises that the behaviour of higher-order (biological) entities, such as man, might both be explanatorily irreducible to (i.e. emergent from) and yet entirely consistent with lower-order (physical) laws.

It is clearly in the human sciences that the propaedeutic work of philosophy is likely to be most rewarding—if only by allowing a better contrast to be drawn between the conditions and possibilities of the natural and social sciences. Thus the non-availability of spontaneously occurring and the impossibility of experimentally establishing closed systems means that criteria for the rational assessment and development of theories in the human sciences cannot be predictive and so must be exclusively explanatory. Again, the concept-, activity- and space-time-dependence of social structures means that any social science must incorporate a historically situated hermeneutics; while the condition that social science is a part of its own field of inquiry means that it must be self-reflexive, critical and totalising in a way in which natural science is not. (See BHASKAR, 1979.)

But transcendental realism has implications for the practice of natural science itself. For it follows from my argument that scientists, when they are engaged in experimental and theoretical work, are implicitly acting on transcendental realism. But it does not follow that they *realise* they are. Nor does it follow that transcendental realism is the only, or even (at any moment of time) the dominant, philosophy they are acting on. One is therefore as a philosopher of science fully entitled to criticise the practice of any science for its lack of scientificity. The importance of this should be clear. For, for example, instrumentalism may be used to impede attempts to build realistic scientific theories, just as empirical realism may be used more generally to suppress alternatives. Of course, the possibility of a realistic description or explanation of any particular level of reality may be bounded in practice by semi-permanent conceptual or technical (or even economic) problems, or by the domain assumptions of the particular science, or by the fact that reality is itself bounded for us there. These possibilities limit internal, but do not refute metaphysical realism. For metaphysical realism says nothing about how much there is to know, or about how much of what there is to know can actually be known by men.

Three main positions characterise the history of philosophical reflection on the natural sciences. For empiricism, the natural order is what is given in experience; for idealism, it is what we make or construct; for

realism, it is given as a presupposition of our causal investigations of nature, but our knowledge of it is socially and laboriously constructed—with the cognitive resources at our disposal, on the basis of the effects of those investigations. For realism, it is the nature of objects that determines their cognitive possibilities for us; it is man that is the contingent phenomenon in nature and knowledge that is, on a cosmic scale, so to speak, accidental.

In science man comes to know man-independent nature, fallibly and variously. This knowledge-relation is both the theme of philosophical reflection and a topic for scientific investigation. But only transcendental realism by setting man *in* nature is consistent with the historical emergence, and causal investigation, of science (or philosophy) itself. Now any such investigation will itself already presuppose an intransitive (and so non-human) ontology of transfactually active and potent structures. This ontology is realism. And it is a necessary presupposition of natural science. But it remains an open question how far, and with what results this principle will actually be carried out in the laboratories and classrooms, journals and monographs, colloquia and conference halls of our actual historical sciences.

References

- BACHELARD, G., 1936, *La Dialectique de la Durée* (London)
- BACHELARD, G., 1953, *Le matérialisme rationnel* (Paris)
- BHASKAR, R., 1975, *A realist theory of science*, 1st edition (Leeds, 1975), 2nd edition (Harvester Press, Sussex, UK and Humanities Press, N. J. USA, 1978)
- BHASKAR, R., 1979, *The possibility of naturalism* (Harvester Press, Sussex, UK and Humanities Press, N. J., USA)
- DRETSKE, F., 1969, *Seeing and knowing* (London)
- ENGELS, F., and K. MARX, 1970, *The German ideology*, ed. C. Arthur (London)
- HANSON, N. R., 1965, *Patterns of discovery* (Cambridge)
- HARRÉ, R., 1970, *The principles of scientific thinking* (London)
- HARRÉ, R., 1973, *Surrogates for necessity*, Mind, pp. 358–380.
- HEGEL, G. W. F., 1949, *The phenomenology of mind* (London)
- HEGEL, G. W. F., 1965, *The science of logic* (London)
- HEMPEL, C. G., 1960, *The theoretician's dilemma*, in: *Aspects of scientific explanation* (New York), chap. 8
- HESSE, M. B., 1974, *The structure of scientific inferences* (London)
- HUME, D., 1948, *Dialogues concerning natural religion* (New York)
- KANT, I., 1972, *Critique of judgement* (New York)
- KUHN, T. S., 1970 a, *The structure of scientific revolutions*, 2nd edition (Chicago)

- KUHN, T. S., 1970 b, *Reflections on my critics*, in: Criticism and the growth of knowledge, eds. I. Lakatos and A. Musgrave (Cambridge), pp. 264–265
- LOCKE, J., 1959, *Essay concerning human understanding*, Epistle to the reader (New York)
- MILL, J. S., 1961, *A system of logic* (London)
- POPPER, K., 1972, *Objective knowledge* (Oxford)
- PUTNAM, H., 1978, *Realism and reason*, Meaning and the Moral Sciences (London), pp. 123–140
- WITTGENSTEIN, L., 1961, *Tractatus logico-philosophicus* (London)

POSSIBLE WORLDS AND THE ONTOLOGY OF A SCIENTIFIC THEORY

V. N. KOSTIOUK

Moscow Institute of Electrotechnology, U.S.S.R.

In my talk I shall consider epistemological problems only, omitting logical considerations. The main problem concerns the ontology of a scientific theory. In this connection I also wish to outline some problems for scientific realism.

The problem of the ontology of a scientific theory is the problem of the existence of postulated entities. The naive view on this problem affirms that those entities exist indeed and just as they have been described in an accepted theory.

The understanding of the unacceptability of this point of view and the comprehension of the falsity of the opposite one has led to have "middle" conceptions in accordance with which the common and traditional philosophical imaginations about reality are erroneous. As W. Sellars says, "speaking as a philosopher, I am quite prepared to say that the common sense world of physical objects in space and time is unreal—that is, that there are no such things" (SELLARS, 1963, p. 173).

But fortunately there is a different image of the world, the so-called *scientific* one. This image is objective, because it is not anthropocentric; it is the true objective story about the real world. As Sellars says to have good reason to accept a scientific theory is to have good reason to believe that the entities it postulates are real.

And contemporary philosopher must change all traditional (for example positivistic) conceptions of reality into some suitable scientific image of the world. This image is a true story about the world in some sense of "truth".

Analogical views were developed by J. SMART in *Philosophy and scientific realism*. As he says, "It is better, however, to face reality and see the world true, as it is". According to his remark, those philosophical views "may correctly be described as materialistic".

Those views developed by W. Sellars and J. Smart include two connected but different fundamental problems. The first one is connected with the clear recognition that there is an objective reality which exists outside and independently of human beings. This line of scientific realism was expressed clearly by R. Tumela and J. Niiniluoto. "Our (critical) scientific realism claims that science is about a reality that exists independently of observers. In principle this reality is knowable though in a symbolic and distorted way. Knowledge about this reality is primarily obtained by means of scientific theorizing and experimentation, and this knowledge always remains corrigible. The important point about this kind of factual knowledge is that it need not be totally empirical... It follows that a realist may conceive of scientific theories as (genuinely) true or false...".

This problem is the most uncontentious in the theory of scientific realism. It is well enough known within the framework of dialectical materialism, and I do not want to discuss it here.

The second problem is connected with an elucidation of what that objective reality concretely is. What do we have in mind when we say that some entities are real? How may we coordinate this assertion with the statement that our ideas about what entities are real change constantly? This is, I believe, the main problem and the main difficulty for different versions of scientific realism.

In particular, I have some doubts about the solution of this problem given by Sellars and Smart. In the first place, there is no essential difference between "common sense" and "scientific" images of the world from the point of view of eternity. Both the first and the second are our ideas about reality, and both after all may be mistaken. Any concrete scientific picture of the world will become absolute sooner or later.

In the second place Sellars and Smart exaggerate the lack of anthropocentricity in a scientific picture of the world, of course, scientific knowledge is impersonal and it has no anthropocentricity in this sense. But any knowledge is produced by a human being, that is why it has some degree of anthropocentricity. In any human knowledge of reality there always arises a question about the attitudes of human beings towards this reality. The anthropocentricity of scientific knowledge is inevitable in this sense.

In this connection a well-known critique of Kant's philosophy by Smart

fails. The latter wrote: "Kant's so-called Copernican revolution was really an anti-Copernican contra-revolution. Just when man was being taken away from the centre of things by the astronomers, and when he was soon to be put in his biological place by the theory of evolution, Kant was by means of this metaphysics, putting him back in the centre again" (see SMART, 1963, p. 151).

But the Kantian problem is something different. He does not try to put any man into the centre of things, he tries to make clear the attitude of man towards things. And every philosopher must do that, too.

The decisive point in this problem is the *separating* of the philosophical definition of objective reality from different concrete "scientific pictures of the world", and the recognition of the deep inner *connection* between them. This was done by V. I. Lenin in *Materialism and empiriocriticism*.

The definition of objective reality is a "pure" assertion of the objective (that is, independent of a human consciousness) existence where there is no indication of objects, properties and relations. This definition says nothing about *what* exists. If we want to know what objects are real, we must use some concrete "scientific picture of the world". The philosophical definition of objective reality may be supplemented by different concrete "scientific pictures of the world". (AMBARCUMJAN and KAZJUTINSKII, 1970).

One may construct a lot of different such scientific pictures of the world. We may talk, for example, about the "physical picture of the world" in a narrow or a broad sense.

Concrete scientific pictures of worlds include some information about objects, properties, connections and relations. But those pictures are *our* ones, that is why we *cannot infer* from them the conclusion that the world is "out there" and independent of us. Such a statement must be constructed *independently* of any scientific picture of the world. If we do not accept this statement, we cannot know about objective and really existent entities studied by science. In this sense scientific realism must have a scientific picture of the world *plus* the philosophical definition of objective reality.

Thus, there is a close connection between a scientific picture of the world and the notion of objective reality. Without a concrete scientific picture of the world the notion of objective reality contains no concrete knowledge about the world and becomes similar to Kant's "thing in itself". Without the notion of objective reality no "scientific picture of the world" can guarantee the objectivity of science.

There are other connections between scientific pictures of the world and the notion of objective reality. Every scientific picture of the world "gives"

some structure to the objective reality. Entities which exist in one such picture may be absent in another one. Ether, electromagnetic field, and so on are the standard examples of such entities.

That is why the complete answer to the question "what entities put together constitute the objective reality?" raises some difficulties. But if our ideas about real existence change constantly, then how can we have a guarantee of the truth of scientific realism? In order to answer this important question we must turn our attention to Quine's (well-known) paper *Ontological relativity*.

As W. Quine says, it is meaningless to ask "Is this thing real?": the question can meaningfully be asked only relative to some background language, some coordinate system. That is why it makes no sense to say what the objects of a theory are beyond saying how to interpret or reinterpret that theory within another. This is a relativistic thesis in Quine's sense.

As he stresses the ontology of a scientific theory (and the scientific picture of the world) "is indeed doubly relative. Specifying the universe of a theory makes sense only relative to some choice of a manual translation of the one theory into the other". (See QUINE, 1968, p. 205.) Therefore, according to Quine, the ontology of any scientific theory or any scientific picture of the world can never be made clear absolutely, but only relative to some other theory.

How much of this thesis of Quine is important for an ontology of a theory of natural science? The language of such a theory is, as a rule, a mixture of different terms: theoretical, observable, formal and informal. Therefore, when we give an answer to a question like "what is an atom?", we have a translation into the same language, because the object language and background one coincide in this case. In this sense Quine's ontological relativity can become trivial.

But in some other sense the ontological relativity of a scientific theory is very important. When a scientific theory is subjected to experimental verification, it assumes the role of an hypothesis. This entails a double uncertainty. In the first place, one does not know to what extent this theory is true. In the second place, one does not know whatever there are entities in objective reality such as this theory describes.

When accepting some theory we must accept some ideas about the structure of objective reality. We have no absolute picture of reality, but every time we get such a description of reality which is true if an accepted theory is true. In this sense a scientist always deals with an ontological relativity. This is a special translation of Quine's thesis into the language of scientific

realism. In this translation the linguistic problems are changed into epistemic ones.

The principal peculiarity of this translation is an acknowledgment of some duality of realistic ontological description: on one side, this description is the recognition of objective reality and on the second side, it is some concrete scientific picture of the world. The informational role of the second aspect is obvious. The first aspect must provide the objectivity of this scientific picture of the world. The notion of objective reality permits us to say: *our* description of reality is objective, because it is relative to things that exist independently of any description.

The scientific picture of the world is subjective without the notion of objective reality. The notion of objective reality has no concrete content without some scientific picture of the world. In this sense both sides of such ontological description are necessary.

The main problem which arises here consists in the grounds of such ontological description. But it is not a new problem, because any ground for a scientific theory is a ground for its ontological implications.

But a further difficulty arises here. The same scientific theory may have different (and incompatible in fact) ontological implications. The standard example is wave and corpuscular descriptions of quantum objects.

But if a theory has incompatible consequences then must it be false? Or must this consequence relation have been understood in a special sense? I believe the second alternative is the case. It is connected with a transition from classical logic to modal logic. The semantics of such a logic uses the fruitful notion of a "possible world".

This notion has applications not only in modal logic, but in mathematics also. And empirical science using mathematical methods gradually comes to adopt such methods of investigation.

What is a possible world in our specific sense? We may say that possible worlds are different ways of partitioning the same objective reality which occurs if a given theory is true. The different ontological implications of a theory correspond to different possible worlds in this sense.

This allows us to ask a question about the truth of a scientific theory understood in a realistic sense. Owing to the principle of ontological relativity, there is not an absolute system of objects according to which a theory may be considered true or false absolutely. It means that a scientific theory can be neither absolutely true nor absolutely false. But we may say that it may be true *relatively*, that is, true according to some way of describing the objective reality.

The notion of relative (differently approximate, that is, relative and objective together) truth is the notion most congenial to scientific realism. I agree completely with R. Boyd who says: "What the realist requires is a notion of approximate truth which allows us to see the actual history of, say, physics as representing the development of increasingly true stories of what the world (both its observable and its unobservable aspects) is like, by successive refinement and modification of good stories into better ones". (See Boyd, 1976.)

But what is the assertion that a scientific theory is approximately true? One may give the following definition of an approximate truth: a theory is approximately true if and only if it is true (in the usual sense) in at least one of the possible worlds admissible according to some suitable scientific picture at the world.

Accordingly, one may say that one theory *is more true than another* if and only if the first is true in all worlds in which the second is true, but not vice versa. If we construct some technical modifications of the notion of a possible world, then we may say that Einstein's theory is more true than Newton's theory. It means that Einstein's theory is true in every possible world in which Newton's theory is true, but not vice versa.

The notion of possible world may also be used to characterize different "world of images" of investigations. Supporters of different theories may use different "worlds of images" and that may lead to mutual incomprehension between scientists.

In order that one scientist may understand another, he must translate the language of "world of images" of his opponent into his own language. But if Quine's thesis about the impossibility of radical translation is true, then mutual understanding between different scientists cannot be complete. A scientist having functionally different roles (for example, as physicist and as philosopher) may not understand himself. There may be a conflict between his own philosophical and his own physical conceptions.

By analogy with this, one may consider the problem of comparison of different scientific theories. There is the well-known point of view that different scientific theories may be incomparable at the psychological, conceptual and ontological levels. If this point of view is correct, then the changeability of scientific knowledge is incompatible with scientific realism. But I believe that this idea is false.

In the first place, all experimental theories of Modern Times are comparable—on their quantitative consequences. It is essential that different scientific languages of Modern Times contain the same quantitative sub-

language—the arithmetic of natural numbers. This comparability fails to provide mutual understanding among scientists but it is enough for rational reconstruction of the choice of a suitable theory or hypothesis.

In the second place, although as Quine says, it is impossible to radically translate one scientific theory into another, there is a possibility of some of a satisfactory translation. The translation is satisfactory if it can explain why some theory which is being translated must be rejected and some other theory (which is the translation) must be accepted. We do not need a third language for this process.

Let me take the frequently discussed trivial example of the translation of classical physics into relativistic physics. I can understand it being said that this translation is not a radical one: we have quantitative consequences in Newton's sense but we have no Newtonian notions. But this is a satisfactory translation, because it allows us to understand in which cases the Newtonian physics is true and in which cases it is not.

Let me stress that in the other direction such a translation is impossible. It is impossible to translate relativistic physics into classical physics even in an extensional sense. (Otherwise these theories would be extensionally equivalent which is impossible.) We have as a result the possibility of cumulative scientific progress. That is that there exists a succession (may be, a subsuccession) of theories in which every theory provides an extensional translation of the previous ones (but not vice versa). Thus we have a vague answer to the question of how scientific realism is possible in spite of ongoing change of ontological assumptions and accordingly a constant change of scientific theories (there may be some gaps in the succession of scientific theories but (for simplicity) I have not considered such delicate cases).

But one important point remains. This is the problem of the possibility of a convergence of such theories (and their ontological assumptions accordingly) toward some “complete scientific picture” of objective reality. Such a convergence must take place if we accept the assumption about knowable reality. This possibility, however, does not logically follow from the previous arguments and suitable postulates must be formulated independently. I believe these postulates may be confirmed indirectly by the history of science.

To conclude let me enumerate the main features of scientific realism as I understand it.

1. The object of a scientific theory is objective reality which is “given” a structure according to the contents of this theory.

2. This results in some “scientific picture of the world” having some degree of correctness. This is the (objective) degree of confidence with which one may assert that the entities postulated are real.

To put it differently: there is no such thing as an absolute structure of the world known by human beings. If there is an absolute structure of the world we do not know it as a whole, and if we know some global structure, then this structure is not an absolute one (principle of ontological relativity).

3. According to this principle a description of reality from the perspective of scientific realism must have two connected aspects: a “scientific picture of the world” plus the notion of objective reality. The first gives concrete information about the world, the second shows that this information is about things that are out *there* and independent of such information.
4. According to the above points one may say that the scientific picture of the world is an approximate one, and that this picture is “approximately true”.
5. By means of the notion of possible worlds such notions as “approximate truth”, “more true than” can be made clear. It may yield different models of infinite scientific progress, mutual understanding of scientists, and so on.

The statements proposed here may certainly be improved and refined. They are but one tentative formulation of a version of scientific realism.

References

- AMBARCUMJAN, V. A., and V. V. KAZJUTINSKII, 1978, (Амбарцумян, В. А., Казютинский)
Научная революция и прогресс в изучении Вселенной, Вопросы философии, Но. 4
- BOYD, R., 1976, *Approximate thruth and natural necessity*, Journal of Philosophy, vol. 73, p. 634
- NIINILUOTO, J., and T. TUOMELA, 1973, *Theoretical concepts and hypothetico-inductive inference* (Dordrecht, Holland)
- QUINE, W. V., 1968, *Ontological relativity*, Journal of Philosophy, vol. 65
- SELLARS, W., 1963, *Science, perception and reality*, (London, New York)
- SMART, J. J., 1963, *Philosophy and scientific realism* (New York)

DIFFICULTIES FOR REALISM IN THE PHILOSOPHY OF SCIENCE

J. J. C. SMART

The Australian National University, Canberra, Australia

In this paper I wish to discuss some difficulties for realism in the philosophy of science.¹ I do this as one who has been concerned to defend realism about the sub-atomic entities of physics (and indeed other unobservables too, such as space-time points). What is such a realism? It is the theory that takes 'There are electrons,' for example, at face value, and neither tries to translate it away into statements about our sense experiences or about macroscopic material things, nor treats it merely as part of a useful instrument for deduction of observable facts. The arguments against translatability² or reduction of the former sorts are too well known to need recapitulation here, and so I shall therefore take the opponent of realism to be

¹ This paper was written while I was a Fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, California. I should like to thank various friends who have commented on an earlier draft of this paper: Nancy Cartwright, David Cole, John Etchemendy, Antony Flew, Bas van Fraassen, Susan Haack, Ian Hacking, Bruce Kuklick, Ruth Barcan Marcus, Wesley Salmon, Robert Stalnaker and Mark Wilson. After completing the first draft of this paper I was fortunate to be able to read a typescript of BAS VAN FRAASSEN'S (forthcoming) book *The Scientific Image*, in which van Fraassen defends a novel and sophisticated form of anti-realism. Properly to discuss van Fraassen's book would require a separate paper. Even before reading his book I had come some way nearer his position: see my tentative suggestions about quantum mechanics in the final paragraph of my paper. However, I still want there to be a realistically interpretable (though perhaps unknown) theory somewhere in the background.

² Apart from more traditional attempts at showing translatability, I also rule out meta-linguistic detours, as in Craig's method (CRAIG, 1956). Even assuming that physics could be effectively axiomatized so as to allow for an application of Craigian transcription—for some scepticism on this score see LEEDS (1975)—the manner in which the transcribed theory is obtained can be understood only by meta-linguistic discourse about an ostensibly realist theory. If we did not take the ostensibly realist theory at face value the success of the transcribed theory would be a mystery.

some sort of instrumentalist. The crudest form of instrumentalist takes talk about electrons to be simply a meaningless instrument of calculation, rather like the beads on an abacus.

The realist's most general argument against instrumentalism is that he can explain the success of microphysics in predicting facts on the macro-level. The instrumentalist just has to accept this success as brute fact. Why should things happen on the macro-level just *as if* there were electrons, neutrinos, and so on, if there really is no micro-level and if discourse that appears to be about unobservable micro-entities is not to be taken at face value? The scientific realist holds that his opponent is left with something like a cosmic miracle. That theories work, or that certain generalizations on the observational level hold true, is something for which his instrumentalist opponent can give no explanation.

Now if this is the realist's argument, the instrumentalist can raise the following objection. To explain the facts on the macro-level by reference to supposed facts on the micro-level is to treat the former as non-accidental relative to the latter. But will the latter facts not be just as much cosmic coincidences themselves? That all electrons are attracted by protons is no less a merely contingent constant conjunction than is some universal and projectible generalization on the macro-level. The law about electrons and protons is expressible in the purely extensional notation of predicate logic and there is no place for a full-blooded notion of natural necessity. The scientific realist can be made to look like John Locke's 'Indian', who said that the earth was supported by an elephant, which was supported by a tortoise, which was supported by... See Locke's *Essay concerning human understanding*, Bk. II, ch. 13, § 19. In his *Dialogues on natural religion*, Section 4, David Hume evidently alludes to this passage in Locke. The comparison with natural theology can be carried further, since van Fraassen, in a delightfully witty article (VAN FRAASSEN, 1974), has compared the sort of scientific realist argument that I have been considering with Aquinas' Third Way.

Van Fraassen follows HICK (1973)³ in re-interpreting the Third Way. That some particular fact is as it is may be explained by saying that the world is thus or so. But that the world is thus or so is equally a contingent fact. A theologian may contemplate this mass of contingency and may even be led to ask why there is a world rather than nothing. Similarly the scientific realist looks at the facts on the observational level, and asks why

³ Hick himself refers (p. 21) to various contemporary Thomist philosophers.

they are as they are. The answer "It is just a matter of fact" does not satisfy him. The theologian is not satisfied with an accidental world, and the scientific realist is not satisfied with an accidental observation level. Aquinas, in his Third Way, concluded to a first cause, God, who was in some sense necessary. Admittedly Aquinas' notion of necessity was not that of a *logically* necessary being, as was the case with Leibniz and Samuel Clarke, but it was that of a being who by his own nature is not subject to decay or dissolution.⁴ Nevertheless the notion of *necessary* non-decay and non-dissolution is hardly more intelligible than that of logically necessary existence, at least if the necessity is thought of as logical necessity, and if the theologian's notion is not that of logical necessity he has not succeeded in explaining clearly what other sort of notion of necessity it is supposed to be. If existence can never be necessary the existence of God would be no less a 'brute fact' than the existence of the universe itself, and so we might as well accept the existence of the universe as ultimate, without trying to go one step further back, with a loss of ontological economy, to the existence of God. Similarly it may be urged against the scientific realist that he is merely replacing one cosmic coincidence by another, with a similar loss in ontological economy.

How as a scientific realist, can I reply to this challenge? Aquinas' argument depends on a notion of necessary existence which I find unintelligible. As a scientific realist I must not fall back on a similarly unintelligible notion of physical necessity. I think that Hume was right in saying that we can have no such (objective) notion of necessity. My reply to the challenge is that there are accidental generalities and accidental generalities. (Compare the view that all men are equal but some are more equal to others.⁵) There are good cosmic coincidences and bad ones, and the bad ones need to be explained by the good ones. That is, by postulating unobservable particles, and so on, and by stating a relatively small number of laws pertaining to these, a scientist can explain the untidy and multifarious facts about the macro-level in a relatively simple and unified manner. In the words of PERRIN (1920, p. vii), we "explain the complications of the visible in terms of invisible simplicity". Not only will the laws on the micro-level be fewer than those on the macro-level, but they will not have to be hedged about by qualifications or *ceteris paribus* conditions, at least to anything like as great an extent.

Here I am using simplicity as a criterion of metaphysical believability: it is not a matter of mere convenience or of saving labour of thought. Of

⁴ See FLEW (1976), p. 55 and BROWN (1964).

⁵ As in George Orwell's *Animal Farm*.

course, we have to accept the macroscopic facts anyway, but if they are seen to follow from simpler principles, then they become less puzzling. Somehow we think that simple theories are *antecedently* more likely to be true than are complex and messy theories. I do not know how to justify this assumption (the assumption "Expect nature to be simple") any more than I can justify the Humean principle "Expect the future to be like the past". The assumption is one that scientists themselves feel to be a natural one.⁶

If I have to rely thus on a general appeal to the above mentioned sort of simplicity, I am of course open to an objection from the side of the instrumentalist that I have assessed simplicity wrongly. The instrumentalist may say that his theory is simpler, because he posits fewer sorts of entities than the realist does.

It would be an advantage, therefore, if the realist could give up his reliance on mere appeal to simplicity (or other aesthetic qualities) and could show that his argument is one of a sort that scientists unquestionably use when arguing to previously unknown facts about the macro-realm, and that it would be unreasonable for them to reject such an argument when it is used in order to assert the real existence of micro-entities. The sort of argument I have in mind is that which has been called 'argument to the best explanation'. But in what way is the realist's explanation the 'best' one? As I have noted, the instrumentalist might be unimpressed by the realist's appeal to simplicity. This could be either because he is not interested in simplicity as such or because he has a different idea of what makes a theory simple. He does not wish to go beyond the regularities, such as they are, which he finds on the macro-level, though he is happy to use merely instrumentally understood theories as a way of connecting these regularities and of predicting new ones.

Dissatisfied, perhaps, with the vagueness of appeal to 'the best explanation', Wesley Salmon has argued that good explanation typically postulates a *causal mechanism*. In SALMON (1978) he adds to his previous account of explanation in terms of a set of 'statistically relevant factors' and 'pertinent probability values' a requirement of 'causal explanations of the relevance relations'. Subsumption under statistical regularities is important, he says,

⁶ For some purposes a justification of the criterion of simplicity may be possible. See SOBER (1975). Nevertheless Sober's justification of simplicity makes use of this criterion itself. Though he holds that simpler theories are in a certain sense more likely to be true (*ibid.*, p. 168) he nevertheless holds that "support and simplicity are irreducibly different goals in hypothesis choice". Sober holds also (*ibid.*, p. 175) that as a methodology realism scores over antirealism as being more simple.

but that "if the regularity invoked is not a causal regularity, then a causal explanation of the regularity must be made part of the explanation of the event" (SALMON, 1978, p. 699). Salmon gives an interesting discussion in its own right of certain forms of statistical and causal explanation; in particular he supplements Reichenbach's notion of a 'causal fork' by means of the notion of an 'interactive fork', thus bringing statistical ideas to bear on the analysis of causality. We postulate a common cause of different events when we find a conjunction of phenomena which would be antecedently improbable, but which are much less improbable (though the probability can be less than 1/2 and even quite small) relative to the hypothesis of a common cause. According to Salmon, this postulation of a common cause is an essential feature of scientific explanation. This may be doubted. Consider, for example, explanations in the theory of relativity which are not causal but geometrical, or explanations in quantum mechanics, such as by reference to the Pauli exclusion principle, which are quite notoriously hard to reconcile with causality.

Still, Salmon could still say that explanation by reference to a causal mechanism is a common enough feature of scientific explanation, and that the instrumentalist is being simply arbitrary when in certain cases he refuses to accept the real existence of unobservable particles. He points out the sort of astonishing coincidence that seems to be explicable only by the real existence of an underlying causal mechanism. Consider first the experimental determination of Avogadro's number with the help of considerations of the kinetic theory of gases and of the theory of the Brownian motion of particles suspended in a gas. Consider secondly the determination of Avogadro's number by measurement of the amount of silver deposited on the cathode during electrolysis of a solution of a salt of silver, and with knowledge of the amount of electric charge required to deposit a single silver ion, which comes from Millikan's and J. J. Thomson's experimental determinations of the charge of an electron. Here we have two very different ways of determining Avogadro's number. The two very different sets of experiment yield the same number within the limits of experimental error. If we do not think realistically about the gas as made up of molecules, of the silver ions lacking one electron each, and so on, why should we expect this antecedently unlikely equality?⁷

⁷ Jean Perrin used such coincidences from at least a dozen different ways of determining Avogadro's number as a clinching argument for the reality of atoms and molecules (PERRIN, 1920, pp. 206-7). See also NYE (1972, p. 171). I am indebted to Wesley Salmon for referring me to these two books.

Surely, says Salmon, this coincidence can be explained only by the principle of the common cause: there must be a common causal mechanism. Perhaps Salmon is using ‘common cause’ a bit widely here, because there is no single event in question here as a ‘common cause’ and to talk of a common *sort* of mechanism here is perhaps stretching things a bit. Moreover, Avogadro’s number is not an event and so cannot have a cause. But whether or not Salmon’s example connects up very readily with his account of explanation as causal, the example is certainly a very striking one, and shows the sort of consideration that many of us find quite persuasive for realism. The antecedent probability of the numerical coincidence from the two sets of experiments is low, but becomes high relative to the hypothesis of electrons, molecules, ions, etc. The instrumentalist does not have a good explanation of this coincidence because he is precluded from regarding his theoretical sentences as *true*. Admittedly if a sentence $\Gamma p \vdash$ is a good computational device a working scientist will probably be willing also to say that $\Gamma p \vdash$ is true, and to assent to the Tarskian paradigm ‘ $\Gamma p \vdash$ is true if and only if p ’. Now this is all right for mere substitutional quantification over ‘ p ’, provided that certain constraints to avoid antinomies are met. But to get the Tarski theory of truth we need objectual quantification over individuals and the notion of ‘satisfaction’. We need things like “(y) ‘ x is an electron’ is satisfied by y if and only if y is an electron”, quantifying over a universe that contains electrons. For the semantics to be genuinely explanatory ‘is satisfied by’ has to be an extensional context, like ‘kicked’ and unlike ‘is a fictional description of’. For Tom to kick the football both Tom and the football must exist.

It is true, however, that the instrumentalist *could* make use of Tarski’s notion of *truth in a model*. We cannot say that a sentence $\Gamma p \vdash$ is true in a certain model M if and only if p , unless of course M just is the *universe*, with the right mappings between the predicates of the language in question and sets of things (or sets of sequences of things) in the universe. If the instrumentalist’s theory is a consistent set of sentences, as he hopes it is, then it has a mathematical model, indeed a model in the domain of the natural numbers. But this model-theoretic fact does not entitle the instrumentalist to say that he is talking about numbers when he *seems* to be talking about electrons and protons. We must not confuse truth in a model with truth *simpliciter*. Moreover even if we did allow the instrumentalist to say that he was talking about numbers and not about electrons and protons, we could accuse him of straining at a physical gnat and swallowing a Platonic camel. I myself want to swallow both the gnat and the camel, for reasons

that are due to Quine: in physics we find sentences quantifying over numbers and sets no less than over electrons and protons. I do feel uneasy about the camel, but can unfortunately see no way of avoiding the hypothesis of Platonistic entities. Indeed in the final paragraph of this paper I shall make some compromise with the model-theoretic sort of instrumentalism, so that in the context of micro-physics I shall indeed myself strain at the gnat and swallow the camel. But if I strain at the gnat I shall want to keep *some* hold on realism in a more indirect way.

To say that realism is the right account of theoretical physics is therefore much the same as to say that the sentences of theoretical physics are true in the objectual Tarski sense. (True, not just true in a model.) So another threat to realism might come from neo-verificationist theories of truth, such as have been proposed by Michael Dummett. Dummett extends an intuitionist way of looking at mathematics to language generally. I must say that I find intuitionism philosophically obscure. Even though the intuitionist may not exactly confuse use and mention when he equates $\Box p \Box$ with ' $\Box p \Box$ is provable', nevertheless the equation of the sense of an object language sentence with that of a metalanguage sentence seems very odd. Nor am I attracted to any form of verificationism generally. However I am not familiar with Dummett's recent theories about all this and would not be able to do justice to them: I shall therefore merely note them as a possible threat to realism.

I now come to another sort of threat to realism, which arises from the fact of scientific revolutions. I have connected realism to truth, but what confidence have we that any scientific theory is true? The Newtonian theory of gravitation, so it may be said, was overturned by Einstein's general theory of relativity, while the Newtonian laws of force and the laws of classical electro-magnetism were replaced by quantum theory. And so on. Making an induction from the revolutions in the past to the probability of revolutions in the future, we may be tempted to say that existing physical theories are almost certainly false. Yet they are quite successful. So it looks as if the realist's argument, the argument from success to truth, must be rejected. The next move is therefore to see whether we can replace the argument from success to truth by an argument from success to *approximate* truth. We might not be able to conclude that there is any object that exactly satisfies the predicate 'is an electron', for example, but we might be able to reassure ourselves that there must be some object that *approximately* satisfies it. That a theory of approximation to truth, or of 'verisimilitude', was required for the defence of realism was clearly stated some years ago

by Sir Karl Popper (see POPPER, 1963, p. 235, 1972, pp. 47–61, 101–103 and 331–335). Popper's qualitative definitions of verisimilitude have been criticized by David Miller, Pavel Tichý, and others, including Popper himself.⁸ In particular see MILLER (1974a) and TICHÝ (1974). For a defence of realism it would be good, in particular, if we could say in what way theories might approximate ontologically to the truth (and hence to one another). Can a theory about particles be *ontologically* an approximation to a theory about fields, for example?

At least, it may be said, there is a sense in which the laws of an old theory approximate to the laws of a new theory. According to PUTNAM (1965) to say that a law or theory is approximately true is to say that a certain logical consequence of it is true *simpliciter*. For example, a correct theory might contain an inverse 2.001 law and a theory that approximated to it might contain an inverse square law. However, what we should like would be a theory of verisimilitude according to which we could say that the new theory, even if not true, is *nearer* the truth than the old theory. MILLER (1975) has argued that unless the laws of theory *B* are exactly true, then even if some laws of *B* are more nearly correct than those of *A* there will be other laws of *B* which are less correct than corresponding laws of *A*. This parallels the criticism of Popper's qualitative theory of verisimilitude in MILLER (1974a), where it is argued that on Popper's theory *B* can be nearer the truth than *A* only if *B* is true *simpliciter*. Perhaps for a merely metaphysical defence of realism we might wish to rely on an account of verisimilitude in terms of mere nearness to an unknown quite-correct theory, but such a notion will be found by anti-realists to be quite vague, or even obscure, if it does not allow us to say, for example, that special relativity is nearer the truth than Newtonian mechanics. We might have wildly differing estimates of the population of China, with no idea which were more correct, but at least we could say what it was for one to be nearer the truth than another. This sort of assurance is lacking in the case of verisimilitude of theories.

Theories that succeed one another in the manner of a Kuhnian revolution can differ from one another ontologically and not just in respect of laws, as when, for example, a field theory replaces a particle theory. For simplicity I shall illustrate the point not by any real scientific theories but by means of an imaginary example. Consider two theories each of which has a sort

⁸ C. E. Mortensen is at present exploring the possibility that if a theory had a basis of relevance logic it might be possible to avoid Miller's objections to an account of its verisimilitude on Popper's lines. Also other accounts of verisimilitude are possible, and further research is in progress (HILPENEN, 1976, NIINILUOTO, 1979, MILLER, 1979).

of pre-Socratic simplicity. Imagine first a cosmology according to which everything is made up of variously shaped solid massy particles, like Democritus' atoms, and let this be replaced by a cosmology according to which instead of particles in a void we have a plenum with variously shaped holes in it. Suppose that these holes move about and rebound from one another much in the way in which the original massy atoms were supposed to do. It is clear that these holes could still be referred to by the same word 'particle'. The adherent of the new theory would of course reject certain sentences of the old theory in which the word 'particle' occurred, for example sentences to the effect that particles were continuously filled with matter, but many sentences in the old theory, such that certain sorts of particles move in ellipses, might still occur in the new theory. Thus a lot of sentences common to the two theories could still contain the word 'particle', and for acceptance of these sentences it would not matter whether particles were thought of as massy lumps in a void or as empty spaces in a plenum. In a sense the two theories would be 'approximately about' the same things. (Nor, of course, would it matter if the new theorists coined a neologism to replace the old word 'particle'.) This sort of case is not too worrying, because there is a one-one correlation between the fundamental entities of the two theories (or at least there would be if the putatively correlated entities existed). A worry is that there may not be such a one-one correlation between the fundamental entities of different theories. Hence the sense in which the ontological statements of successive theories approximate to truth is still very obscure.

Some of the sentences that the rival theorists will use will be observation sentences, which can be characterized, following QUINE (1974, pp. 39–41), as those sentences which would be assented to (or dissented from) by all members of a linguistic community who were in the appropriate perceptual situations. Thus a true observation sentence might be to the effect that the needle of a certain ostensively defined device pointed to the numeral '1.5'. Those who knew that the device was a properly functioning ammeter would also of course assent to the statement that the electric current through the device was 1.5 ampères. On the other hand, many members of the linguistic community would not understand the expression 'ampère' or 'ammeter' and so the sentence 'The current through the device is 1.5 ampères' would not count as an observation sentence, though it could be made to do so if we restricted the linguistic community to scientifically trained persons who knew that the ammeter was correctly calibrated and so on. Now if we think of observation sentences in this way we can see that even revolutionary

changes in physics will be unlikely to affect observation sentences. Even if a revolutionary theory led us to say quite novel things about electrons and electric currents it would not lead us to give up our assent to sentences such as that the current through a circuit was 1.5 ampères.

Nor indeed need such a revolutionary change of our theory about electrons lead us to change our assent or dissent to many sentences about electrons themselves (i.e. containing the word 'electron'), such as that, for example, electrons and protons have opposite electric charge. Whatever revolutions there are in the theory of the nucleus, and whatever hypotheses may be developed to explain the present complicated taxonomy of elementary particles, we can be sure that water will still be H_2O and that a hydrogen atom will still contain one proton and one electron. We must not allow the fact of scientific revolutions to blind us to the solid mass of ascertained and never-to-be overturned theory in physics. 'Hydrogen' and 'electron' are not likely to go the way of 'phlogiston'. I conclude therefore that provided a good theory of verisimilitude can be worked out (and this is a big 'if') realism has probably nothing to fear from the fact of scientific revolutions.

Approximation to truth may be thought of in terms of reference to 'ideal entities'. But it is easier to talk in terms of 'ideal entities' than it is to know what one is talking about. Are the 'real entities' of current theory, such as electrons and protons, no more than 'idealizations', like the continuous fluids of hydrodynamics, the isolated dynamical systems of classical mechanics, or the isolated harmonic oscillator whereby students are introduced to quantum mechanics? There is something confusing in saying that physics talks about idealizations, since idealizations do not exist in the real world, and so are not there for the theory to be about, in a referentially transparent sense of 'about'. There is an intensional sense of 'about', in which we say that Dickens wrote a book about Mr. Pickwick or that Mallory wrote one about the knights of the Round Table, but I think we should analyze this as purporting, pretending, or something of the sort, to 'talk about' in the extensional sense; i.e. not in terms of actually operating an intensional semantics but in terms of purporting (etc.) to operate an extensional one. (In a sense I want to say that there can be no intensional semantics.)

Now the problem of ideal objects in science is more difficult to deal with than is that of outright fiction. Neither Dickens nor Mallory were providing explanations, and their stories have no use in prediction. The explanatory and predictive efficacy of 'ideal theories' needs to be explained. One possible

explanation may come from the predicates of the ideal theory being approximately satisfied by real objects (or sequences of them) and not from their being exactly satisfied by ideal objects (or sequences of them).

In conclusion, however, I want to canvas a possible explanation of the explanatory and predictive value of physics which is a compromise between an out-and-out realism and instrumentalism. I do not think that this compromise should have been used prior to the development of quantum mechanics, and perhaps it can be put aside in the future if quantum mechanics is replaced by some more satisfactory theory. Many of us do, however, sense a special difficulty in giving a completely realistic philosophy of quantum mechanics.

Earlier in this paper I argued that realism was plausible, because it gave the best explanation of the predictive success of physics and because only if there is a realistic theory can we avoid supposing an implausible cosmic coincidence on the observational level. But plausibilities must be weighed relatively to one another, and there are things about quantum mechanics that make a plausible realistic interpretation hard to achieve. Consider the sort of situation envisaged in EINSTEIN, PODOLSKI and ROSEN (1935), in which there are two interacting systems S_1 and S_2 , which then separate. By doing appropriate experiments on S_1 it is possible to determine properties of S_2 . Einstein, Podolski and Rosen used this as an argument for realism, because alternative experiments on a system S_1 could make it possible to determine at will either the position or the momentum of S_2 , or perhaps one or other of some other pair of conjugate properties. Since S_1 and S_2 could by this time be distant and not interacting, this suggests that S_2 had these two conjugate properties all along, contrary to the uncertainty principle. Subsequent investigation showed that such a realistic interpretation was not possible, and indeed the argument now goes against realism. It appears that an experiment on S_1 does actually change the properties of S_2 , even though the two are distant from one another.⁹ This seems to smack either of subjectivism or of action at a distance, and both alternatives are unpalatable. (Of the two, however, I would choose action at a distance, if I had to choose.)

If we think that a realist interpretation of quantum mechanics is not possible, must we go back to a thoroughgoing instrumentalism or some other form of anti-realism?¹⁰ A realist interpretation of quantum mech-

⁹ The evidence has been surveyed and assessed in CLAUSER and SHIMONY (1978).

¹⁰ Such as van Fraassen's model-theoretic approach, in his forthcoming book mentioned in footnote 1.

anics lacks plausibility, but so on the other hand, as I have urged earlier in the paper, does an anti-realist metaphysics of science. I think that perhaps we can get a compromise which implies an ultimate metaphysical realism but which allows an instrumentalist view of present-day microphysics. Let us recall the fact that the instrumentalist might deny straight-out truth to theoretical sentences but can talk of their truth in a model. A consistent set of sentences must have a mathematical model. Now at the stage of classical physics we shall have come to postulate a domain of mathematical entities no less than of physical particles, since (as Quine has urged) physics is tested holistically and contains mathematics as an integral part of it. We can surely retain this ontology of mathematical entities when we pass to microphysics, and so we can believe realistically in a domain of mathematical entities that provides a model for the sentences of microphysics. Let us regard these sentences of microphysics as purely instrumental, but nevertheless we may suppose that the success of this instrumentalist theory may perhaps be explained by the idea that in some appropriate sense the model of the instrumentalist theory is an approximation to some mathematical model of some unknown but true realist theory. This sort of instrumentalism would therefore be realist in principle. Not only would it be realist about mathematical objects (which is not important in the present connection) but it would point to an unknown realist theory in the background. I do not know whether this sort of compromise between realism and antirealism could be worked out. In particular, I do not know whether this sort of model-theoretic situation could obtain without the supposedly 'instrumentalist' theory approximating to the realist theory in the background, i.e. possessing verisimilitude, and so bringing us back to an earlier attempt at defending realism.

References

- BROWN, Patterson, 1964, *St. Thomas on necessary being*, Philosophical Review, vol. 73 pp. 76-90
- CLAUSER, John F., and Abner SHIMONY, 1978, *Bell's theorem: Experimental tests and implications*, Reports on Progress in Physics, vol. 41, pp. 1881-1927
- CRAIG, William, 1956, *Replacement of auxiliary expressions*, Philosophical Review, vol. 65 pp. 38-55
- EINSTEIN, A., B. PODOLSKY, and N. ROSEN, 1935, *Can quantum-mechanical description of reality be complete?* Physical Review, vol. 47, pp. 777-780
- FLEW, Antony, 1976 *The presumption of atheism* (Pemberton Publishing Co., London)
- HICK, J. H., 1973, *Philosophy of religion* (Prentice-Hall, Englewood Cliffs, N. J.)

- HILPENEN, R., 1976, *Approximate truth and truthlikeness*, in: M. Przelecki, K. Szaniawski, and R. Wójcicki (eds), *Formal Methods in the Methodology of Empirical Sciences* (D. Reidel and Polish Scientific Publishers, Dordrecht and Warsaw), pp. 19–41
- LEEDS, Stephen, 1975, *A note on Craigian instrumentalism*, *Journal of Philosophy*, vol. 72, pp. 177–184
- MILLER, David, 1974a, *Popper's qualitative theory of verisimilitude*, *British Journal for the Philosophy of Science*, vol. 25, pp. 166–177
- MILLER, David, 1974b, *On the comparison of false theories by their bases*, *British Journal for the Philosophy of Science*, vol. 25, pp. 177–188
- MILLER, David, 1975, *The accuracy of predictions*, *Synthese*, vol. 30, pp. 159–191
- MILLER, David, 1976, *Verisimilitude redeflated*, *British Journal for the Philosophy of Science*, vol. 27, pp. 363–381
- MILLER, David, 1979, *On distance from the truth as a true distance*, in: Jaakko Hintikka, Ilkka Niiniluoto and Esa Saarinen (eds), *Essays on Mathematical and Philosophical Logic* (D. Reidel, Dordrecht), pp. 415–435
- NIINILUOTO, Ilkka, 1979, *Truthlikeness in first order languages*, in: Jaakko Hintikka, Ilkka Niiniluoto and Esa Saarinen (eds), *Essays on Mathematical and Philosophical Logic* (D. Reidel, Dordrecht), pp. 437–458
- NYE, Mary Jo, 1972, *Molecular reality* (Macdonald, London)
- PERRIN, Jean, 1920, *Atoms* (Constable, London)
- POPPER, K. R., 1963, *Conjectures and refutations* (Routledge and Kegan Paul, London)
- POPPER, K. R., 1972, *Objective knowledge* (Clarendon Press, Oxford)
- PUTNAM, Hilary, 1965, *How not to talk about meaning*, *Boston Studies in the Philosophy of Science*, vol. 2, pp. 205–222. Reprinted in: *Philosophical Papers*, vol. 2 (Cambridge University Press), pp. 117–131
- QUINE, W. V., 1974, *The roots of reference* (Open Court, La Salle, Illinois)
- SALMON, Wesely C., 1978, *Why ask, 'Why?' An Inquiry Concerning Scientific Explanation*, *Proceedings and Addresses of the American Philosophical Association*, vol. 51, pp. 683–705
- SOBER, Elliott, 1975, *Simplicity* (Clarendon Press, Oxford)
- TICHÝ, Pavel, 1974, *On Popper's definitions of verisimilitude*, *British Journal for the Philosophy of Science*, vol. 25, pp. 155–160
- VAN FRAASSEN, B. C., 1974, *Theoretical entities: The five ways*, *Philosophia*, vol. 4, pp. 95–109

METATHEORETICAL DILEMMAS OF THE SOCIAL SCIENCES

THE CASE OF SOCIOLOGY

PIOTR SZTOMPKA

Jagiellonian University, Kraków, Poland

*The clash of doctrines is not
a disaster but an opportunity*
(Alfred N. Whitehead)

1. Critique of sociology; two traditions

Judging by the level of critical self-consciousness, the social sciences are the most developed disciplines indeed. Perhaps in no other branch of science an equal amount of attention has been paid to its own deficiencies and failures, and no other branch of science has spent so much of its creative potential for self-destructive purposes. Criticism and self-criticism is certainly indispensable for the progress of knowledge, but different forms of criticism are by no means equally fruitful in this respect. If one reflects on the paradox that in spite of the overabundant criticism, the social sciences are still far from scientific maturity, one of the possible explanations may be found precisely in the predominance of the less fruitful forms of criticism, to the neglect of more promising ones. In this paper I shall focus on the discipline of sociology, but the discussion will also be relevant for other social-scientific disciplines.

Two alternative emphases are possible if one is looking critically at the sociological enterprise. First, there is the *surface-level* of problems, methods, results, and applications. All this can be grasped directly, by external observation of what is going on in the discipline; what are the questions most often posed, what are the research-ways most often followed, what are the results most often yielded, and what are the uses most often made of those results. As a rule, the careful scrutiny leads to the most pessimistic appraisals, the source of failures is located in the youth, immaturity, retardedness of sociological research, and the proposed therapy comes down, accordingly, to the piecemeal, fragmentary improvements;

a sort of gradual "coming of age" of sociology as a science. I shall refer to this approach as the *poverty-of-sociology* trend in the criticism of the discipline.

Second, there is the *depth-level* of underlying assumptions; premises of the methodological, epistemological, and ontological type which are logically, necessarily presupposed by the given practice of the discipline, whether its practitioners are conscious of that or not. All this can be grasped only indirectly, by interpretation or reconstruction of what is assumed, or rather—what must have been assumed if the actual course of research and the actual content of results are to have any meaning. If the depth-level of assumptions rather than the surface-level of actual practice becomes the target of criticism, the message is equally often pessimistic; sociology is said to remain in the pre-paradigmatic stage, to lack any adequate theoretical approach, etc. But here the similarity ends. The source of defects is no longer located in the immaturity of the discipline, but rather in the fundamentally mistaken nature of initial premises. And the proposed therapy is no longer restricted to the "reformist" policy of piecemeal, fragmentary improvements, but rather calls for the "revolutionary" overthrow of existing paradigms, a total re-orientation of the research-practice and the introduction of a more viable and adequate approach. The criticism of this sort has become the fashion of the day only recently, and is usually referred to as the *crisis-of-sociology* trend.

By now my personal bias is I think obvious, but let me state it for the record. I believe that the poverty-of-sociology tradition is relatively unproductive and fruitless. Lamentations about the outrageous sins of our discipline usually lead nowhere, and the policy of letting sociology to mature, would require sociologists to wait for another century or more, before applying for a scientific status. I am not that patient, for one. On the other hand, the crisis-of-sociology approach seems to me most promising and fruitful. The debate focusing on the essential, fundamental assumptions of sociology opens up an opportunity for a basic reorientation and progress.

2. Polar assumptions; two sociologies

What is wrong with the theoretical assumptions accepted by "academic sociology", "conventional sociology", "mainstream sociology", "traditional sociology"—as the critics usually define our discipline? Several diagnoses

are presented, but the leading theme seems to be the plurality and polarization of assumptions. The plurality of standpoints in itself is rarely regarded as something wrong, basically detrimental to the further development of sociology; just the reverse, the metaphor of the "marketplace of ideas" has always positive connotations. But when the plurality is accompanied by the strict separation, dogmatic closure, or polarization of standpoints, with no possible mutual openings—the critical concern is born. When the "marketplace of ideas" changes into the "battlefield of ideas"—science usually exits by the back door.

The situation in which the typical solutions of the fundamental meta-theoretical questions are phrased as mutually opposite and exclusive, in "either... or..." terms is most characteristic for sociology. The *meta-theoretical dilemmas* are generated by almost any question that can be asked of the methodological, epistemological or ontological roots of the discipline. Let me enumerate those which seem most crucial.

The first *methodological* problem encountered by any theoretically-oriented sociologist deals with the relationship of his own domain of inquiry to other similar domains. This may be labelled as the problem of demarcation. In sociology, this problem gives birth to the quest for the precise delimitation of the place and role of our discipline in the realm of sciences. In the debate on this issue two more specific questions arise: (A) what is the relationship of sociology to the natural sciences, and (B) what is the relationship of sociology to psychology? Responding to the question (A) sociologists face the dilemma: science or humanities? Perceiving sociology as basically similar to the natural sciences, they accept the *naturalistic* assumption, and seeing it as basically different from the natural sciences, they accept the *antinaturalistic* assumption. Responding to the question (B) sociologists face another dilemma: science of man, or science of society? Perceiving sociology as secondary and subservient to psychology they accept the *reductionist* position, and conversely—defining it as an autonomous and independent discipline they accept the *antireductionistic* standpoint.

The second, *epistemological* problem of immediate relevance for any sociologist attempting to produce a theory, has to do with the nature of those ultimate theoretical results which can be obtained. The dispute tends to concentrate on two more specific questions:

- (A) what are the functions expected of a theory?, and
- (B) what is the structure of a theory; what types of propositions it allows, and what types it forbids.

Responding to question (A), sociologists encounter the dilemma: knowledge or action? Those who restrict the functions of a theory to the provision of adequate explanations, and consequently allow only categorical propositions (as opposed to prescriptive, or normative ones) in the body of a sociological theory, presuppose the *cognitivistic* assumption. On the other hand, those who consider a theory mainly in terms of its practical impact, and consequently require it to offer normative, prescriptive advice for action, presuppose the *activistic* position. Responding to question (B), sociologists come upon the dilemma: detachment or bias? Some refuse any place for valuations or value-judgments in the research process and the resulting theories. They profess the totally objectivistic attitude, free from any valuations as the only fruitful approach to sociological data. In this way they endorse the *neutralistic* assumption. On the other hand, those who accept valuations as the necessary component of sociological research, and value-judgments as the necessary ingredient in sociological theories, endorse the opposite, *axiologistic* standpoint.

The third, *ontological* problem is perhaps the most significant of all. It deals with the fundamental properties of the subject matter studied by sociology. The quest for the substance and scope of social facts generates two more specific issues:

- (A) what is the nature of man, the ultimate component of society, and
- (B) what is the nature of social wholes (groups, collectivities, institutions, civilizations, etc.).

Responding to question (A), sociologists face the dilemma: man as an object, or as a subject? Those who conceive of a human being as totally molded, determined, overpowered by the external influences (societal or otherwise) accept the *passivistic* assumption. On the other hand, those who perceive man as inner-directed, self-controlling, independent (at least to some degree) of external influences, accept the *autonomistic* assumption. Responding to question (B), sociologists face the dilemma: society as a whole or an aggregate? Some consider society as a simple sum of individuals (their activities, plus eventually—the results of such activities). Thus they endorse the *individualistic* assumption. Others ascribe to society some sort of super-individual existence, positing specific properties and specific regularities that pertain exclusively to the social wholes. Thereby they subscribe to the *collectivistic* assumption.

Those six metatheoretical dilemmas, which I consider as most salient in sociology, are not mutually independent. There exist definite, systematic interlinks among assumptions; some of them tend to cluster together,

some tend to be mutually exclusive. The typical clusters of assumptions taken in the response to basic methodological, epistemological and ontological questions will be referred to as the *models of sociology* (methodological, epistemological and ontological, respectively). The modified classification of assumptions may be presented, taking into account their typical, mutual interlinks.

(A) Methodological models of sociology

- (a) Scientific (including naturalism + reductionism, as essential components).
- (b) Humanistic (including antinaturalism + antireductionism as essential components).

(B) Epistemological models of sociology

- (a) Objectivistic (including cognitivism + neutralism as essential components).
- (b) Critical (including activism + axiology as essential components).

(C) Ontological models of sociology

- (a) Mechanistic (including passivism + collectivism as essential components).
- (b) Voluntaristic (including autonomism + individualism as essential components).

Now, it may be observed that the regular relationships may also obtain between the specific models of sociology. Usually, the accepted ontology is linked with the accepted epistemology, and this in turn—with the accepted methodology. Two comprehensive clusters of this sort may finally be distinguished; two master models, or perhaps "paradigms" of sociology. One includes the scientific methodology, objectivistic epistemology, and mechanistic ontology. It will be labelled as the *positivistic model of sociology*. The other includes the humanistic methodology, critical epistemology, and voluntaristic ontology. It will be labelled as the *subjectivistic model of sociology*. By the logic of my analytic argument those models of sociology are mutually contradictory and mutually exclusive. They represent the polar ideal types of the prevailing modes of sociological inquiry. Their opposition is responsible for the current crisis of sociology.

3. Sociologists' reactions to ambivalence

The acute sense of a crisis, the anomie situation in which the sociologists found themselves without the guidance of any unambiguous, and universally accepted paradigm, led to several characteristic responses. To draw a parallel with Merton's famous study of anomie and normlessness,

one may describe those typical patterns as ritualism, retreatism, and rebellion.

In the case of the *ritualistic pattern*, the sociologists follow blindly the routine research-ways fashionable in their days, without any consideration given to the validity or utility of their final results. Research becomes a self-justifying enterprise, quite independently of the fruit it bears (or most often—does not bear).

In the case of the *retreatist pattern*, a certain suspension of theoretical ambitions, at least for the time being, is recommended, and theoretical as well as methodological eclecticism is raised to a position of virtue. There is only one step from here to the anti-methodology of “anything goes” and theoretical anarchism.

The opposite response to a crisis of sociology may be identified as the *rebellious pattern*. Here, in place of the traditional ways of science, the totally new goals as well as new means are suggested. The only trouble is that one is no longer sure whether what is at stake is still science. One may point, on the one hand to such schools as “ethnomethodology” or “phenomenology” within the more general trend labelled as the “new humanism”, and on the other hand, to “radical sociology” or “reflexive sociology”—within the other general trend which may be labelled as the “new activism”. With the multiplication of closed, exclusive and often dogmatic schools, this type of reaction simply adds up new dimensions to the theoretical chaos, and practically aggravates the crisis.

Neither of the above reactions therefore seems satisfactory. The response to the crisis of sociology which is to my mind most promising could be labelled as *innovative pattern*. I accept without any qualifications the traditional goal of science—constructing the explanatory and predictive theory. But I reject the traditional approaches to such a goal, both those typical for positivistic sociology, and those typical for subjectivistic sociology. I believe that the multiple dilemmas responsible for the crisis of sociology must be overcome, and in this process new solutions devised for the fundamental, metatheoretical issues of sociology. Only in this way can the full creative potential of the crisis be utilized, and the crisis itself—surmounted.

4. The strategy of dialectic criticism

The method I propose to use in order to overcome the traditional dilemmas may be called the *dialectic critique of assumptions*. In each case I shall attempt to formulate a third, alternative solution, which would

save what is valuable ("the rational core") of both extreme standpoints, but at the same time reach a qualitatively new level of theoretical insight. The thesis embedded in one of the extreme assumptions will be combined with the antithesis embedded in the opposite assumption, to produce a *dialectic synthesis*, in the form of a third, alternative solution. In this way, each of the traditional dilemmas will be shown to be spurious, and will be overcome.

To be more specific, the method proceeds in the following fashion. If the whole dilemma, with both extremes, is to be rejected, and replaced with a new solution, then a new solution will obviously have to be opposite to both extreme positions. By the same token, it will have to be opposite to everything those extreme positions have in common. Thus the first step is to discover a set of assumptions of the higher-order shared by both extreme positions within a traditional dilemma. Then, the second step is to negate those higher-order assumptions. And finally, the third step is to find out what solution to the original question which generated the traditional dilemma is entailed by the new set of higher-order assumptions. This solution will represent the dialectic synthesis, and will become a building-block for the new model of sociology, namely *dialectic sociology*.

5. An illustration; the riddle of values

Let me illustrate the dialectic strategy with one example. I will choose the epistemological dilemma of *neutralism* and *axiology*, dealing with the place of valuations in the research process and the research results of science.

Is there anything in common between the axiologists and the neutralists, accepting and rejecting valuations, respectively? At the first glance, in the extreme ideal-typical formulations outlined earlier, both standpoints are completely opposite, and mutually exclusive. But if one looks under the surface and uncovers the hidden higher-order presuppositions of neutralism and axiology, an important area of commonality in the similar approach to the problem of valuations will be discovered.

I claim that both standpoints accept in fact the same notion of scientific objectivity which shall be further labelled as the *traditional notion of objectivity* (or *objectivity₁*). This notion can be explicated by means of four higher-order assumptions. First, the objectivity is an absolute property of scientific research and/or scientific results, i.e. it can be predicated of research or of results per se, by critically examining their internal form

or structure. Second, the property in question, defining objectivity in the absolute fashion is the lack of bias; or to put it in other words—objectivity is the converse of bias. Third, a bias is always the result of valuations, and valuations always produce bias. Fourth, valuations are therefore contradictory to objectivity; valuational activities preclude the objectivity of scientific method, and the presence of value-judgments precludes the objectivity of scientific results. Thus, according to this conception, if any valuational activities are spotted in the scientific research, its method can for this very reason be judged as non-objective; and if any value-judgments are spotted in the scientific knowledge, such a knowledge can, for this very reason, be treated as non-objective, too.

Traditional notion of objectivity (objectivity_1) is accepted by the neutralists and axiologists alike. The neutralists hold that objectivity_1 is attainable in the social sciences and should be strived for. The axiologists hold that objectivity_1 is unattainable in the social sciences and should be rejected as the unrealistic and improper goal. The dialectical way to overcome the dilemma of neutralism and axiologism must lead through the rejection of such notion of objectivity, and replacing it with some alternative notion. I shall call it the *new notion of objectivity* (or objectivity_2). It may be explicated by means of the four composite assumptions, direct opposites of the previously listed ones.

First, objectivity is now conceived as the relative property of a scientific research and/or scientific results, i.e. it cannot be predicated of research or results *per se*, by inspecting its form or structure. Second, the relativization has two-fold nature: (a) Objectivity of results is relative to the real state of affairs in the given domain of reality. It is tantamount to the semantic relationship of adequacy or correspondence between the content of a scientific proposition and the phenomena, processes, or events as they really occur. To put it briefly; scientific results are objective if and only if they are true. Objectivity (of results) is the synonym of truth. (b) Objectivity of research-process is relative to the truth of results achieved by means of research. It is tantamount to the pragmatic relationship of instrumentality between the scientific method and the scientific knowledge. To put it briefly: the scientific method is objective if and only if it is instrumental for the attainment of the true scientific knowledge. Objectivity of a method is synonymous with its reliability, or fruitfullness in the search for true knowledge. Third, value judgments do not impair the truth of research results, and valuations in the research process do not prevent the attainment of true results. They do not produce bias, or at

least they do not have to produce bias. Rather, the contrary is true—value judgments are the necessary, indispensable ingredients of the true social knowledge, and valuations are the necessary, indispensable activities in the procedure rendering true knowledge. Fourth, therefore value judgments and valuations are not the obstacles but rather the prerequisites for scientific objectivity.

The insights suggesting a new notion of objectivity have been occurring to several authors, particularly those who took the natural-scientific methodological patterns to be improper for the social sciences, and struggled for the creation of a new, antinaturalistic, and antipositivistic methodology, suitable for the study of man and society. More generally, the intellectual tendency to move from the traditional toward a new notion of objectivity is, I believe, a characteristic trait of the contemporary social sciences.

Let me trace some implications of the new notion of objectivity for the initial problem: the place and role of valuations in science. The most obvious one is the breaking of correspondence (or identity) between the concept of value involvement and the concept of bias. In the traditional notion there existed a definitional link between the two. Now, each may be treated as an independent dimension.

If the *aspect of valuations* is put into focus, two opposite standpoints may now be defined as:

AXIOLOGISM₁ = sociology cannot and should not avoid valuations.
NEUTRALISM₁ = sociology can and should avoid valuations.

On the other hand, if the focus is on the *aspect of bias*, a different pair of standpoints appears:

AXIOLOGISM₂ = sociology cannot attain unbiased knowledge and utilize unbiased methods.
NEUTRALISM₂ = sociology can attain unbiased knowledge and utilize unbiased methods.

As both dimensions are independent, those two dichotomies may be cross-combined, without the danger of logical inconsistency.

| | | The aspect of valuations | |
|-----------------------|-------------------------|--------------------------|-------------------------|
| | | AXIOLOGISM ₁ | NEUTRALISM ₁ |
| The aspect of bias | NEUTRALISM ₂ | COMMITMENT | OBJECTIVISM |
| | AXIOLOGISM ₂ | SUBJECTIVISM | |

Four possible solutions to the problem of valuations are produced. Three of them are particularly interesting, the fourth being an analytic possibility with no recognized referent in the sociological practice. The combination of AXIOLOGISM₁ and NEUTRALISM₂ renders a new standpoint, not encountered in the traditional debate. I shall call it the standpoint of *commitment*. It is clearly informed by the new notion of objectivity, and it claims that sociology can and should employ valuations as the necessary ingredient of an unbiased method leading to the unbiased knowledge of social reality.

The standpoint of commitment satisfies my criteria for the viable dialectical alternative, overcoming the traditional dilemma of valuations. First, because it rejects, at least in part, both traditional positions. Against extreme neutralism it holds that sociology cannot and should not be value-free. Against extreme axiologism it argues that sociology need not abandon its claims for the true knowledge and unbiased method. Second, because at the same time it provides a continuity, at least in part, with respect to both traditional positions. With extreme neutralism it holds that sociology can be unbiased. With extreme axiologism it argues that sociology cannot escape valuations. Third, which is clearly entailed by the foregoing—it is essentially a new standpoint, not identical with any of the traditional ones.

6. Toward a dialectic sociology

Similar, dialectic strategy can be applied to all six of the traditional metatheoretical dilemmas. The scope of the present paper does not allow to go into details. But if the dialectic solutions to each of the six dilemmas are combined together, a coherent system of assumptions appears, dealing with the status of sociology, its goals, its reach, and the nature of its subject-matter. It may be considered a new *dialectic model* of sociology.

In the dialectic model, as constructed here, sociology is conceived as a distinct science of society, following the general rules of the scientific logic, but within those rules developing a specific methodology suitable for the study of its peculiar subject-matter: man and society and history. This solution to the dilemma of naturalism and antinaturalism may be labelled as the *integralist* standpoint. Sociology is further conceived as a science linked with the other sciences of man by mutual reductive relationships, with a provision for a significant residuum of irreducible concepts and laws, to be included in specific sociological explanations.

This solution to the dilemma of reductionism and antireductionism will be referred to as the *separatist* standpoint. Sociology is also seen as directly relevant for social practice, and social practice as providing an ultimate corroboration of sociological knowledge. Thus a dilemma of cognitivism and activism is replaced with the synthetic standpoint of *constructivism*. Sociology is further seen as approximating the objectively adequate knowledge, due to the self-conscious adherence to specific values, and implementation of those values in the research process, and the body of research results. Thus, instead of neutralism and axiology, the standpoint of *commitment* is proclaimed. Men, the ultimate components of society are perceived as collectively creative and productive, within the scope of opportunities and limitations set by the social and non-social environment, as well as by the historical heritage. In place of passivism and autonomism, the assumption of *creativism* appears. And finally, society—the proper subject-matter of sociology—is conceived as a plurality of individuals bound together by a specific structure of inter-individual relationships, and consequently displaying some emergent properties and regularities of its own. The traditional dilemma of individualism and collectivism is replaced with the assumption of *structuralism*.

This is a definite image of sociology, basically at odds with both traditional alternatives—the positivistic sociology and the subjectivistic sociology. In my view, the further development of sociology must follow along the lines suggested by a dialectic model if sociology is to realize its calling: to become a science *of* society, but also a practical force *for* improving society.

Note: The extended version of the argument outlined in this paper, as well as the full bibliography, can be found in my book: *Sociological dilemmas: Toward a dialectic paradigm* (Academic Press, New York 1979).

PROBABILITY: THE DIFFERENT VIEWS AND TERMINOLOGIES IN A CRITICAL ANALYSIS

BRUNO DE FINETTI

Rome, Italy

1. Probability: what is it about?

It would be improper to ask at once “what is probability?”; the key problem is to specify, first of all, *what probability is about*.

Seemingly, there is no doubt at all: probability is a property concerned with “events”. Yes, but, what is an “event”?

It is from the answer to this question that the main and opposite conceptions arise, leading to very different interpretations of the notion of probability, of its foundations and applications and of the properly corresponding terminology.

According to the most common terminology—borrowed from statistical jargon—the term “event” is meant as a collective term, and every single instance is denoted by “a trial of such an event”. This way, it usually appears understood (and it is really so) that all such “trials” are supposed to be equally likely (and also, perhaps, “independent”); if not, it will usually be specified that “probability is changing from one trial to another” and/or that “the trials are not independent, but positively or negatively (or partly positively and partly negatively) correlated”.

In order to avoid such ambiguities in terminology and all the dangerous misconceptions and misinterpretations ensuing from them, the natural way out, the needed Ariadne’s Thread, is the terminological improvement of calling each *single trial* or instance an “event”.

(If one should wish to use a term for *event* in the rejected *collective meaning*, one could use, for example, “phenomenon”, saying “an event which is a trial of a given phenomenon”. However, when “event” is meant

in the sense of a single trial and all possible specifications may be added, such wording is inessential.)

Analogously, it seems proper to call a "random quantity"—say X —an unknown quantity, instead of "random variable" according to the usual jargon. To say that X has as possible values x_1, x_2, \dots, x_n , each with its probability p_1, p_2, \dots, p_n (or, in general, has a probability distribution $F(x)$, maybe also a density $f(x)$) does not mean that X should "vary", assuming all such values, but only that its true value is known to be one of the x_i with probability p_i .

Such simple terminological precautions (and analogous ones, with the same aim in the same spirit) should be sufficient to avoid ambiguities when speaking about probabilities and when thinking and reasoning in terms of probabilities.

2. "Statistical" misconceptions

The worst cause of obscurities and misunderstandings in the field of probability is the tendency of many authors to identify—or "almost to identify"—probability and frequency.

It is true that, between these two notions, there are many strict relations, but it is just because of this that it is essential to avoid any confusion. If not, we would be inadvertently but unavoidably entangled in a satirical comedy like the ones in which two identical twins are continually mistaken the one for the other.

In order to unify the notations for *events* and for *random quantities* (the usual wording "random variable" is improper since nothing is "varying" but only more or less *unknown*), it is very useful to *identify* any event E with the number usually called its "indicator": $E = 1$ if E is true and $E = 0$ if not. This way, for instance, $X = E_1 + E_2 + \dots + E_n$ is the "number of successes" among n such events, X/n is the frequency, and so on (irrespective of the fact that the values of the E_n and X are known or unknown). Analogously, in general, a random quantity (with a finite number of possible values) may be written

$$X = x_1 E_1 + x_2 E_2 + \dots + x_n E_n.$$

It is useful (at least in my experience) to use the same symbol, \mathbf{P} , for both *Probability* and *Prevision*, where *Prevision* is a substitute for *Expectation*, giving us the (useful) chance to identify the initial letters, and reminding us that probability is but a particular case of prevision.

The worst misconceptions are the ones in which a small probability is counted as impossibility and a large one as certainty. There are strange examples of such attitudes. Most of them are related to regularity or irregularity; several authors (for example Richard von Mises) maintain that every random sequence *must* be "irregular"; for him this is an *axiom*: the "Regellosigkeitsaxiom".

In a different context, Emile Borel expressed the opinion that "a very small probability, e.g. 10^{-20} , is equivalent to impossibility". It seems, at any rate, that such questions are troubling, sometimes even for great men.

Misunderstanding at such a banal level often arises, e.g. by wondering whether an event having a very small probability has occurred. Any event, if strictly specified, is (or was) very improbable. This is what some logicians called "the lottery paradox": anybody having a lottery ticket has a very small probability of winning the big lot, but notwithstanding this someone will be the winner. This is not paradoxical, but obviously natural; it seems paradoxical when irregularity is considered to be a distinctive feature of sequences created by "chance". So it is equally not paradoxical if the big lot is won by a man having the name Hannibal (the only man with this name in his country); however, most people would consider this, not only a curious but also a contradiction of the principle of large numbers.

3. Independence and exchangeability

To mention here the notion of *exchangeability* should not be interpreted as presenting or discussing such a notion: what seems interesting and productive of clarification is to compare, as for terminological property, this word, "exchangeability", with the denotation sometimes employed of "independence, with constant but unknown probability".

However, such wording is contradictory: the compatibility of such conditions cannot hold except in the standard case of a known composition. If, on the contrary, this composition is unknown to us, the result of every trial modifies (according to the Bayes rule) our opinion about the possible composition of the urn. The effect is, of course, that, among the possible hypotheses about the composition of the urn, the ones in better agreement with the observed frequency acquire progressively a greater weight in the "mixture". (An exchangeability scheme is, indeed, a mixture of the elementary schemes of "repeated trials", independent and with a constant probability.)

Roughly speaking, one could say that there is no substantial difference between our conclusion and the current ones; however, the conceptual difference is radical since all the logical aspects of this process (of probability forecasting) are viewed as occurring *in our minds*, whilst the real process develops as it does: not compelled to respect prescribed "regularities" but often doing so, because such "regularity" is, in a sense, "natural" under the usual conditions.

The remark about exchangeability is but a particular case of an obvious but usually understood (or ignored?) remark of general validity and essential importance. It is meaningless to speak of the probability of a given event E , say $\mathbf{P}(E)$ "tout-court" the correct and complete form is $\mathbf{P}(E|H_0)$: probability of E when our state of information is summarized in H_0 . That means that every probability is a conditional one; conditional on our present state of information, H_0 ; if also it is properly conditional (conditional on another *uncertain* event, H), it should be written $\mathbf{P}(E|H_0 H)$, or $\mathbf{P}(E|H)$, only if the H_0 may be understood without ambiguity (from the context or from the nature of the question).

4. Final remarks

The examples discussed are very simple, but it seems to me that, in the field of probability, the major and peculiar difficulties, the most possible causes of misunderstanding, are to be found in vagueness and in the terminology rather than mathematical ones.

For more detailed discussion of the subject it could be illuminating to compare the paper by Hamaker (ISI Review, 1978) with my answer (which appeared in the same journal).

Rome, June 27, 1979

PARADOXES OF CONGLOMERABILITY AND FIDUCIAL INFERENCE*

T. SEIDENFELD

Department of Philosophy, University of Pittsburgh, Pittsburgh, Pennsylvania, U.S.A.

1. Introduction

The issues I propose investigating in this paper trace their origins to the long standing project of giving a precise probabilistic representation of rational belief states and, in particular, to the problems in representing belief states that presystematically would be described as states of near *ignorance*. No doubt, Laplace's principle of Insufficient Reason is the most familiar and the most general attempted solution. He proposed, as part of the definition of credal probability, that where all one knows is a finite partition of 'equi'-possibilities, $\{s_1, \dots, s_n\}$, pairwise exclusive and mutually exhaustive, this state of near *ignorance* is represented by a credal probability that assigns equal values ($1/n$) to each possibility. Thus, in the absence of evidence relevant to the assessment of these alternatives, ignorance goes hand-in-hand with symmetry of credal probabilities. (Two objections to Laplace's principle will be the focus of our analysis in subsequent sections.) It is this trait linking ignorance with symmetry of credal probability that identifies offspring of Laplace's principle, offspring bred to resist the infirmities plaguing Insufficient Reason. I have in mind, particularly, the theory of invariants (a part of H. JEFFREYS' (1967), Objective Bayesianism, structural inference (due to D. A. S. FRASER, 1968), and fiducial probability (due to R. A. FISHER, 1973). (This lineage is least evident in fiducial theory and most evident in Objective Bayesianism¹.) There is a second and, I believe, equally

* Preparation of this paper was undertaken, with support from the Learning Research and Development Center of the University of Pittsburgh, as a Buhl Fellow.

¹ See LINDLEY (1958) for clarification of the tie connecting fiducial probability and Insufficient Reason.

important family tie among these three programs: the common strategy of solving inverse inference by reduction to direct inference. (This lineage is least evident in Objective Bayesianism and most evident in fiducial theory².)

What I offer here is a sketch of a reconstruction of fiducial probability that emphasizes the reduction of inverse to direct inference while providing responses to several objections that purportedly apply as a result of its other heritage: as a descendant of Insufficient Reason. Specifically, using this reconstruction, I hope to respond to the objections raised by Professor STONE in his paper (1979). In this regard, fiducial probability is a heuristic device, a guide through the maze of the 'marginalization paradoxes'. But much more than a guide it cannot be, for the combination of ancestries (Insufficient Reason and the reduction of inverse to direct inference) is lethal. That is, I respectfully submit, the three inductive programs: Objective Bayesianism, structural inference, and fiducial probability, are non-viable hybrids. But, it is equally important to recognize that the fatal flaw common to these programs is not, contrary to what one expects having heard Professor Stone's arguments, the purported inconsistencies of the marginalization paradoxes stemming from the use of improper distributions. In short, I suggest that the intriguing anomalies constructed by Professor Stone (et al.) are a heterogeneous lot which involve at least two logically separate postulates, neither of which is necessary for fiducial probability (in particular), or for sound inductive logic (in general).

Let me summarize these points. As is highlighted in Professor Stone's marvelously clear first (discrete) example, conglomerability lurks behind some versions of the paradox (if a contradiction is forthcoming). But, de Finetti argued thirty years ago, (denumerable) conglomerability is equivalent to countable additivity³. Following this, recent efforts by

² I have discussed the role of direct inference in Jeffreys' use of invariants in SEIDENFELD (1979 b).

³ See DE FINETTI (1972, p. 99). Conglomerability is (for denumerable partitions) the requirement that: if for each element h_i ($i = 1, \dots$) in a pairwise exclusive and mutually exhaustive partition, the inequality

$$k_1 \leq p_{bk}(H, h_i) \leq k_2$$

is satisfied, then

$$k_1 \leq p_{bk}(H) \leq k_2$$

unconditionally.

de Finetti argues for the equivalence of (denumerable) conglomerability and countable additivity. It appears that his proof from conglomerability to additivity rests on (what DUBINS (1975) calls) the non-remoteness of the probability measure.

DUBINS (1975), HEATH and SUDDERTH (1978), and especially LEVI (1980) offer reconstructions of improper distributions using only finitely additive probability (for Levi, finitely additive credal probability is obtained from σ -finite measures, of which the improper distributions are one variety). Fiducial probability does not require countably additive probability; hence, conglomerability is not valid in fiducial theory.

Not all of the marginalization paradoxes are, in my opinion, problems of mere conglomerability. As in Professor Stone's second interesting example: using the normal distribution with both parameters unknown, the argument depends upon rules for transformation of continuous random variables and rules for extracting marginal distributions from the resulting, transformed probability functions. These rules are open to serious criticism even with countably additive probability, e.g. as Kolmogorov's theory is (see KOLMOGOROV, 1956). Those problems already present in the traditional theory are magnified when finite additivity is introduced through improper distributions. A simple remedy, applicable in the familiar case of countable additivity, apparently resolves the ailment for the more intricate second version of the marginalization paradox⁴.

Last in the triad of conclusions is the claim that, despite success at avoiding the marginalization paradoxes, fiducial inference is unsound because of reliance upon uniform distributions (in location parameter problems) to represent ignorance. Just as Laplace's rule is sensitive to reparameterization, i.e. a non-linear transformation of the parameter destroys uniformity, so too fiducial inference is sensitive to the kind of data that appear in inverse inference. A mere reversal of the order of two kinds of experiments reveals the incoherence of this sensitivity.

2. On direct and inverse inference, and Laplace's rule

Let us begin with a review of the distinction between direct and inverse statistical inference, and a rehearsal of the role the Laplacean principle plays in solving inverse inference. Throughout this presentation I shall use two interpretations of the probability calculus: a credal probability, $p_{bk}(\cdot)$, and a stochastic probability or chance $P_k(\cdot)$. Credal probability, p_{bk} , is to be understood as a measure of rational degrees of belief or as

⁴ I have discussed this matter in SEIDENFELD (1980). With this remedy, one can resolve those paradoxes for fiducial inference that depend upon transformation of continuous variables. I noted these as a special case in SEIDENFELD (1979a, pp. 162-3), but offered no solution there.

logical probability (probability₁, in Carnap's sense): where 'bk' stands for a corpus of background knowledge that is evidence for the rational agent. Stochastic probability, P_K , is to be understood as an objective (agent independent), statistical property of processes such as coin flipping, urn sampling or the measuring of physical quantities with observations subject to random error; where ' K ' stands for the *kind of trial*, the process involved.⁵

In direct inference, inference is *from* knowledge of chances *to* hypotheses about some (particular) outcome of the stochastic process, inference from 'population' to 'sample'. For instance, given the information that this is a *fair* coin when flipped by device F , information that the chances are equal ($= \frac{1}{2}$) for *heads* and *tails* on this kind of flip, direct inference fixes a credal probability of $\frac{1}{2}$ for the hypothesis that the *next* flip will land heads up.⁶ So much is non-controversial. But, this account leaves unanswered the central question in direct inference (an unavoidable question for the reduction of inverse to direct inference): what is the principle of direct inference with respect to particular outcomes about which we know more than is contained in our knowledge of chances? For example, suppose the next flip of this fair coin remains in the air for more than two seconds. What if we do not know the chances on such elongated flips of fair coins? Do we ignore this added fact about the next flip and, thereby, treat the duration of the flip as though it were stochastically irrelevant to flips of a fair coin? As we note shortly, this policy-for direct inference: suppressing data about which we do not know the, chances, is, as a general rule, incoherent.

In inverse inference, inference is from knowledge of particular outcomes to hypotheses about the unknown chances on that kind of trial.⁷ For instance, based on the evidence that the coin was flipped n times, of which m trials landed heads up, what is the credal probability for the hypothesis that the unknown chance of heads (on this kind of trial) is in the interval

⁵ I shall not undertake discussion of attempts, like Professor de Finetti's, that seek the reduction of stochastic probability to concepts involving credal probability only, e.g. in terms of exchangeability and partial-exchangeability, nor shall I discuss Professor Kyburg's program, epistemological probability, that seeks the reduction of stochastic probabilities to observed frequencies (see KYBURG, 1974).

⁶ HACKING (1965) solves such direct inferences by his *frequency principle*.

⁷ Since, for de Finetti, there are no 'unknown chances', we must ask the derivative question of singular predictive inference in order that these problems arise on his account of statistical inference.

$m/n \pm \varepsilon$? There is no non-controversial principle of inverse inference that is the counterpart of the simple direct inference principle. Alternative inductive logics differ most dramatically in their respective solutions to inference from ‘sample’ to ‘population’.

A powerful inductive logic obtains by combining Laplacean Insufficient Reason, the direct inference principle, and the postulate of conditionalization. (Note: Conditionalization is the requirement that conditional probability, $p_{\text{bk}}(h; e)$, the conditional credal probability of h given e , fixes the agent’s commitments for updating the credal probability of hypothesis h , were evidence e added to his corpus of knowledge.) For example, let us assume that the background knowledge for inverse inference about the unknown bias of the coin, inference to be based on a sample of n flips, includes the statistical specification that the kind of trial is binomial, i.e. the chance of m heads on n (independent) flips satisfies the binomial distribution:

$$P_K(m \text{ heads on } n \text{ flips}) = C_m^n \cdot \theta^m \cdot (1-\theta)^{n-m}, \quad [\text{A}]$$

where the parameter θ , $0 \leq \theta \leq 1$, designates the unknown chance of heads, e.g. $\theta = \frac{1}{2}$ corresponds to a *fair* coin. Bayes’ theorem for credal probability entails that:

$$p_{\text{bk}}(h_\theta; o_{(m,n)}) \propto p_{\text{bk}}(o_{(m,n)}; h_\theta) \cdot p_{\text{bk}}(h_\theta), \quad [\text{B}]$$

where h_θ designates a (simple) hypothesis of the unknown chance and $o_{(m,n)}$ an observation report—that m of n flips landed heads-up. Thus, subject to conditionalization, inverse inference is proportional to the product of direct inference and the unconditional credal probability for the hypothesis h_θ . In Bayesian jargon, the posterior probability of h_θ is proportional to the product of its likelihood and prior probability. We agree that the likelihood, $p_{\text{bk}}(o_{(m,n)}; h_\theta)$ as a function of h_θ , is fixed by the direct inference principle. Of course, the prior $p_{\text{bk}}(h_\theta)$ is the credal probability which serves to represent the initial ignorance the agent pleads about the unknown chances. If we adopt Laplace’s principle to represent this ignorance with a uniform probability over possible values of θ , then (assuming the unknown chance of heads may be any quantity in the unit interval $[0, 1]$) the posterior probability determined according to [B] duplicates Bayes’ 1763 solution, a Beta probability distribution:

$$\beta(m, n-m) = \frac{(n+1)!}{m!(n-m)!} \cdot \theta^m (1-\theta)^{n-m}. \quad [\text{C}]$$

(*Note:* This posterior corresponds to a Beta prior $\beta(0, 0)$, and supports Laplace's Rule of Succession.)

As familiar as this argument is, so also are two important objections leveled against it:

1. Some (e.g. Fisher) argue that all credal probabilities must reflect knowledge, not merely ignorance of chances. Thus, it is proper to assign a credal probability of $\frac{1}{2}$ to the hypothesis that the next flip of this *fair* coin lands head-up since the direct inference is based upon knowledge of chances (that the coin is fair), knowledge of statistical facts. But, when Laplace's principle is used to assign equal probability to (equi-)possibilities (as in inverse inference), this assignment is lacking the support of the knowledge of chances. In other words, this criticism notes that Insufficient Reason conflates ignorance with knowledge of chances for, if we posit the coin (of unknown bias) is produced by some minting machine (a hyper-population), the uniform credal 'ignorance' probability over hypotheses about the unknown bias is also the credal probability that would obtain by direct inference if, counter to fact, we knew that the minting process was stochastic, producing coins of differing biases with a uniform chance distribution. That is, Laplace's principle generates ignorance credal probabilities indistinguishable from credal probabilities derived by direct inference. As we see (below), this objection fuels the reduction of inverse to direct inference which is the driving force behind fiducial inference.

2. A second, and more compelling objection to Insufficient Reason is the argument that it breeds inconsistencies. Even in the elementary inverse inference (above), we must provide a privileged partition of the possibilities to avoid contradictions. For example, reparameterize the problem from θ to $\zeta = \ln(\theta/(1-\theta))$ and a uniform credal probability for possible values of ζ (with range the real line plus the limiting endpoints corresponding to the endpoints $\theta = 0$ and $\theta = 1$) is incompatible with the uniform probability over values of the bias. If I understand Professor Stone's position, he would argue that we face problems even were we to resolve the matter of which partition to use. For, if we decide upon the parameterization in ζ (based upon familiar reasons (see LINDLEY, 1965), that the uniform distribution over ζ is the conjugate prior to the Beta family, i.e. is the limiting distribution of proper priors as $a, b \rightarrow -1$ in $\beta(a, b)$), the resulting prior probability distribution is improper and the marginalization paradoxes loom.

If this is an accurate account of Professor Stone's objection to Insufficient Reason (or to its offspring: invariance, structural inference, and fiducial probability), then I do not agree with his diagnosis, since I believe we can explain-away the marginalization anomalies without forgoing these versions of Laplace's program. The reconstruction of fiducial inference I offer (in the next section) demonstrates this possibility. However, I maintain that these theories are unacceptable for a reason closely related to the second objection to Insufficient Reason: that inconsistencies result because of alternative parameterizations.

3. On fiducial inference and the marginalization paradoxes

How does the fiducial argument accomplish the reduction of inverse to direct inference?⁸ So that we might discuss Professor Stone's second illustration of the marginalization paradox (STONE, 1972, p. 370), I review inverse inference for the normal distribution. First, I shall present a one-parameter fiducial argument and then generalize this to several parameters. (Only after completing the full, two parameter inference will we be able to respond to Professor Stone's challenge.)

Suppose we are faced with inverse inference based on measurements subject to normal error. For example, imagine that we use an unbiased scale of known precision to weigh an object about which, otherwise, we are ignorant. Let our background knowledge include the normal stochastic model for observed weights w_i , that is, we assume that the measurement process is modeled by the normal distribution $N(W, \sigma^2)$, where W , the mean of this chance distribution is the true (unknown) weight of the object and σ^2 , the variance of this distribution, is a known constant corresponding to the imprecision of the instrument. Without loss of generality, let $\sigma^2 = 1$. Then, the chances for an outcome w on this kind of trial are given by the normal density:

$$f_K(w) = (1/\sqrt{2\pi}) \exp [-(w - W)^2/2]. \quad [D]$$

In conjunction with conditionalization, before trial, direct inference fixes conditional credal probability for an observation o_w given the unknown weight h_w , i.e. direct inference specifies values of $p_{bk}(o_w; h_w)$. But, also, the stochastic model provides unconditional direct inference about hypothetical random variables: *pivotal* variables. Consider a sample of n (independent) weighings of the object, $\{w_1, \dots, w_n\}$, and let \bar{w} be their

⁸ The central theme of this reconstruction is due to JEFFREYS (1967). Later writers who followed his idea include DEMPSTER (1963a), HACKING (1965), and KYBURG (1974). Hacking's analysis is noteworthy for its lucidity.

average. Then the quantity, $v = (\bar{w} - W)$ has a known chance distribution $N(0, 1/n)$, and (before trial) by direct inference we have precise credal probability for h_v , statements about v , i.e. $p_{bk}(h_v)$ is well defined by direct inference.⁹

What becomes of this simple direct inference about v after the trial, after \bar{w} is observed? Following Jeffreys' idea (JEFFREY, 1967, p. 381) (developed, in distinct ways, by DEMPSTER, 1963a; KYBURG, 1974; and HACKING, 1965), we discover that if we are prepared to treat the extra information about \bar{w} as credally irrelevant to the simple direct inference about v , i.e. if

$$p_{bk}(h_v) = p_{bk}(h_v : o_{\bar{w}}) , \quad [E]$$

the result is precise posterior credal probability about W . In other words, if, given the datum \bar{w} , credal probability about v is summarized by saying that v has an $N(0, 1/n)$ distribution, then, given the datum \bar{w} , credal probability about W is summarized by saying that W has an $N(\bar{w}, 1/n)$. But this is just the inverse probability, $p_{bk}(h_W : o_{\bar{w}})$ we have been seeking. Thus, with assistance of an irrelevance assumption, inverse inference about the unknown weight W is reduced to direct inference about the pivotal variable v .

Unfortunately, the stochastic model provides many pivotal variables: random quantities that are a function of the data and the unknown parameter with known chance distributions. This is unfortunate, because reducing inverse inference about W to direct inference about a pivotal leads to contradictions if the irrelevance assumption applies to all pivots. For instance, the quantity $v' = (w_m - W)$, where w_m is the sample median, is also pivotal with chance distribution $N(0, \pi/2n)$. But the sample of n weighings cannot be irrelevant to both simple direct inferences about v and v' (the result would be contradictory credal probabilities about W).¹⁰

⁹ The argument that the pivotal v has a known chance distribution $N(0, 1/n)$ does *not* suppose additivity with respect to chances. Nor does it fall victim to de Finetti's criticism of disjunctive syllogism for conditional probability (DE FINETTI, 1972, p. 104). Specifically, de Finetti's objection applies to disjunctive syllogisms over different kinds of trials. But the argument for pivotal chances is valid, as it is limited to a single kind of trial, e.g. repeated measurements of the object on this scale. See LEVI (1980, 12. 14), also LEVI (forthcoming) and my, SEIDENFELD (1979a, 9.1, 9.6).

¹⁰ I have used this example to criticize epistemological probability theory for failing to satisfy the sufficiency principle (in inverse inference), (SEIDENFELD, 1979a, pp. 193-5). Professor S. Spielman (City University of New York) has been kind enough to suggest (in correspondence) an interesting rebuttal to this argument. He points out that, in the

(I note, in passing, this much similarity with Professor Stone's (first) discrete example of the marginalization paradox: For that problem there are two pivots—

- v_1 : with two possible values, annihilation/non-annihilation, and
- v_2 : with four possible values, electron/positron/muon/antimuon.

Given the stochastic model (provided by Professor Stone) the pivots have known chance distributions and, prior to observation, by direct inference there are precise (marginal) credal probabilities for the hypotheses h_{v_1} and h_{v_2} . But the sample cannot be irrelevant to all these direct inferences. That is, after trial, we must be prepared to change our credal probabilities for hypotheses about one pivotal, at least. This observation fails to explain the anomalous character of Stone's example. It serves merely to point out that the data must be relevant to one pivotal, at least¹¹.)

The immediate challenge for reconstructing fiducial inference is to provide constraints (of a non-*ad hoc* character) that identify one family of pivots such that the sample data are irrelevant to direct inference about only and any of these related pivots, and such that the resulting fiducial probability is coherent. As Lindley pointed out over twenty years ago (LINDLEY, 1958), if fiducial inference is coherent, it has a Bayesian model; that is, one must be able to provide a Bayesian reconstruction of the fiducial argument complete with a prior credal probability, $p_{bk}(h_\theta)$, that serves to represent the initial ignorance about the unknown quantity θ .

narrower reference class obtained by conditioning on the difference between the sample average and sample median, no conflict in inverse inference about the population mean is present and, moreover, the results that follow are numerically identical to those obtained from the pivotal using the sample average and the original reference class.

However, we may upset this response by showing the new candidate reference class is defeated by Kyburg's rules for randomness. This is easily done by dividing the reference class, based on the difference between the sample statistics, into two (overlapping) classes. Let k_1 and k_2 be two constants chosen to satisfy the inequalities:

$$k_1 \leq m \leq k_2,$$

where m is the sample median (k_1 and k_2 might arise from rounding the observations). Let r_1 be the reference set of samples with median greater than or equal to k_1 . Let r_2 be the reference set of samples with median less than or equal to k_2 . Each of r_1 and r_2 defeats the candidate reference class for randomness about the unknown population mean. Also, r_1 and r_2 defeat each other. The only surviving reference class, $r^* = r_1 \cap r_2$, supports vacuous solutions, with epistemological probability intervals $[\epsilon, 1-\epsilon]$. Hence, as required by the criticism, sufficiency fails since knowledge of the sample median defeats inference based on knowledge of the sample average.

¹¹ According to the reconstruction of fiducial inference offered here, only pivotal v_1 supports the fiducial argument. There is no joint pivotal (v_1, v_2) available.

The following three constraints limit pivots to those that satisfy Lindley's necessary and sufficient conditions for coherence (in one parameter exponential models, with a sufficient statistic):

P_1 —the pivotal must be a function of the parameter of interest and a *sufficient* statistic of the data available (with respect to this parameter).¹² Thus, $v' = (w_m - W)$ is inadmissible;

P_2 —the pivotal must be smoothly invertible (in the sense of TUKEY, 1957).¹³ Thus, v_1 (annihilation/non-annihilation) is inadmissible;

P_3 —the pivotal must be canonical.

(DEFINITION. A pivotal v is *canonical* if the chance distribution of the pivotal is the chance distribution of the sufficient statistic (one of its arguments) for some parameter value θ^* in the parameter space.¹⁴

Examples. $v = (\bar{w} - W)$ is canonical at the value $\theta^* = 0$.

v_2 (electron, ..., antimuon) is canonical for each of the four parameter values compatible with any datum.)

If we examine the supporting Bayesian models for reconstructed fiducial inference, we find (thanks to LINDLEY's (1958) results) that the credal probability representing ignorance (the prior probability of Bayes' theorem) is (for location parameter formulation) the familiar Laplacean uniform distribution. Thus, we expose the connection between fiducial inference and Insufficient Reason. However, if the parameter space is unbounded (as in Stone's discrete example, or as in our problem of weights), the uniform credal distribution is improper.

What is an improper distribution? Its role in the supporting Bayesian model for fiducial inference (and also for orthodox Neyman-Pearson confidence interval theory) is as an 'ignorance' probability distribution. But, formally, an improper distribution is not a credal probability, e.g. the uniform, improper density over the real line is not a probability density. In many problems, the improper distribution that serves as the 'ignorance' prior is conjugate for the statistical model specified by the background

¹² This condition may be weakened to allow exhaustive estimation, as described in SEIDENFELD, 1979a.

¹³ For continuous distributions, a pivotal is smoothly invertible when, for any possible set of observations inserted as arguments, the mapping from parameters to pivotal quantities:

- (I) has the same range for any possible set of observations;
- (II) is 1-1, and hence has a single valued inverse; and
- (III) this inverse is continuous (with continuous derivatives).

¹⁴ Justification for this constraint is offered in SEIDENFELD, 1979a, § 4.4.

evidence, e.g. the uniform density is conjugate for the normal distribution—it is the limit of normal distributions of increasing variance. This explication of impropriety, as the limit of proper distributions, is not fully satisfactory, for it masks the feature of improper distributions crucial to understanding Professor Stone's problem in *linear subatomics*. The limit of countably additive probability distributions need not be countably additive.

Recent work by DUBINS (1975), HEATH and SUDDERTH (1978), and particularly LEVI (1980), follow up earlier analysis by DE FINETTI (1972) in offering somewhat different accounts wherein improper distributions are unmasked for what they are—mathematical representations of finitely additive probabilities. Thus, admitting impropriety (in the form of improper ‘ignorance’ probability functions) carries a commitment for finitely and not countably additive credal probability. De Finetti claimed the equivalence between (denumerable) conglomerability and countable additivity in 1949. STONE's illustration (1976; 1979) of serious impropriety is just the phenomenon of non-conglomerability of finitely additive probability. Our reconstruction of fiducial inference does not require countable additivity. (*Note:* At most we require what Dubins calls π -conglomerability, disintegration with respect to a given margin—but not with respect to arbitrary partitions.) Thus, there is no paradox (in the sense of a contradiction) in asserting:

$$p_{bk}(\text{annihilation}) = \frac{3}{4} \quad \text{and} \quad p_{bk}(\text{annihilation}; o) = \frac{1}{4},$$

where ‘o’ denotes any of the possible experimental outcomes.

Instead of pursuing the counter-intuitive features of finitely additive probability, let us return to the reconstruction of fiducial inference with several parametres. My purpose with this exercise is to demonstrate that Professor Stone's second marginalization paradox does not depend solely upon conglomerability, but upon the mathematics of transformations of continuous variables. Once again, the explication of impropriety in terms of limits of proper distributions masks the crucial point.

Suppose, as before, that we take n (independent) weighings of the object (unknown weight) on the unbiased scale whose errors are normally distributed. However, unlike the earlier version, suppose that the precision of the scale is unknown. The background information specifies the two-parameter $N(W, \sigma^2)$ stochastic model and i.i.d. data (w_1, \dots, w_n) . There exists a pair of (jointly) sufficient statistics (\bar{w}, S^2) and the two-parameter fiducial inference is built up, step-by-step, based on the factorization:

$$P_K(\bar{w}, S^2; W, \sigma^2) = P_K(\bar{w}; W, \sigma^2) \cdot P_K(S^2; \sigma^2). \quad [F]$$

(*Note:* This factorization provides the partition in which π -dis-integration is assumed.)

Step 1: Given σ^2 , \bar{w} is sufficient for W and $v = (\bar{w} - W)$ is a smoothly invertible, canonical pivotal that induces fiducial probability for W , given σ^2 .

Step 2: If S^2 is marginally sufficient for σ^2 (in the absence of knowledge about W), then $v' = (S^2/\sigma^2)$ is a smoothly invertible, canonical pivotal that induces (marginal) fiducial probability for σ^2 .

Step 3: Multiplying the conditional (Step 1) and marginal (Step 2) fiducial probabilities yields the joint fiducial probability. However, to detach the consequent of Step 2, i.e. to grant the assumption that S^2 is marginally sufficient for σ^2 requires a coherence check, verified through the supporting Bayesian model.

Roughly put, each of the first two steps commits one to corresponding ‘ignorance’ prior probabilities: the improper uniform distribution over W and the improper distribution uniform over $\ln\sigma$. Letting the Bayesian model be obtained by taking the product of these priors, we have all the ingredients for verifying the claim that S^2 is marginally sufficient for σ^2 . The reason we are forced to introduce this coherence check (absent in the one parameter case) is the lack of a counterpart concept ‘marginal sufficiency’ to duplicate sufficiency. In the normal distribution, whether or not S^2 is marginally sufficient for σ^2 depends upon how we represent ignorance about the mean W . (The problem of marginal sufficiency is, in this context, equivalent to the worry over marginalization paradoxes for σ^2 .) In fact, the claim of marginal sufficiency is borne out by the coherence check of the Bayesian model generated from the ‘ignorance’ priors, expressed by the improper densities: $dW \cdot d\sigma/\sigma$.¹⁵

The marginal, posterior (fiducial) probability for the quantity:

$$t = \sqrt{n(n-1)} \cdot (W - \bar{w})/S \quad [G]$$

is a ‘Student’s’ t -distribution ($n-1$ d.f.). As we noted above, since this credal probability arises from finitely additive prior probability, failures of conglomerability are to be expected with the t -distribution. As early

¹⁵ The coherence check blocks Dempster’s alternative fiducial argument for this inference problem (DEMPSTER, 1963b).

as 1963, Buehler and Feddersen uncovered one such failure, though its status *vis-à-vis* conglomerability (and therefore its relevance to finitely additive probability) was not noticed (see BUEHLER and FEDDERSEN, 1963).¹⁶

I would like to conclude this discussion of fiducial inference for the normal distribution with a rebuttal to Professor Stone's marginalization paradox for this solution to inverse inference. Following the analysis offered in STONE and DAWID (1972), Professor Stone argues that:

The joint posterior (fiducial) probability for the parameters (W, σ) , is given by the density

$$\sigma^{-(n+1)} \exp[-n(W - \bar{w})^2/2\sigma^2 - S/2\sigma^2] \cdot dW d\sigma. \quad [H]$$

Define the parameter $\theta = W/\sigma$ and transform the posterior probability [H] for the pair (θ, σ) , with density

$$\sigma^{-n} \exp[-n\theta^2/2 + n\theta\bar{w}/\sigma - R^2/2\sigma^2] \cdot d\theta d\sigma, \quad [I]$$

¹⁶ Buehler and Feddersen made the following observations: With two observations, x_1 and x_2 , the t -distribution with 1 d. f. is a Cauchy distribution with quartiles at the two quantities x_1 , x_2 . Hence, for any pair (x_1, x_2) the posterior marginal probability for the unknown mean W has the property that:

$$p_{bk}(h_{(x_{\min} \leq W \leq x_{\max})}; o(x_1, x_2)) = \frac{1}{2}. \quad [L]$$

But, also, for pairs (x_1, x_2) satisfying the inequality

$$|x_1 + x_2| / |x_1 - x_2| \leq \frac{3}{2}, \quad [M]$$

the chance of the two observations straddling the population mean W is not less than .51; hence, by direct inference,

$$p_{bk}(h_{(x_{\min} \leq W \leq x_{\max})}; o_{[M]}) > .51. \quad [N]$$

Equations [L] and [N] are not contradictory unless conglomerability (over a partition with cardinality of the continuum) is used to derive

$$p_{bk}(h_{(x_{\min} \leq W \leq x_{\max})}; o_{[M]}) = \frac{1}{2} \quad [O]$$

from [L].

HEATH and SUDDERTH (1978) show that the posterior probability corresponding to the fiducial solution in this problem is *coherent* in their sense. But, according to their view, the *conditional* bets (taken from the Buehler-Feddersen paradox) are incoherent. Thus, Heath and Sudderth's account of coherence differentiates between *called-off* and *conditional* bets. They do not respect weak-dominance of called-off bets; hence, the called-off bets corresponding to the Buehler-Feddersen paradox are coherent (in their sense).

where $R^2 = \sum w_i^2$. Obtain the marginal posterior probability from [I] by integrating σ out of the joint distribution, yielding the density

$$\exp[-n\theta^2/2] \int_0^\infty \omega^{n-2} \exp[-\omega^2/2 + r\theta\omega] d\omega, \quad [\text{J}]$$

where $r = n\bar{w}/R$.

This density depends upon the statistic r only. Also, the chance distribution (hence, direct inference about) r depends upon θ only, as seen by the density

$$(1 - r^2/n)^{(n-3)/2} \int_0^\infty \omega^{n-1} \exp[-\omega^2/2 + r\theta\omega] d\omega. \quad [\text{K}]$$

But the latter is *not* a factor of the former (as a function of θ). Hence, argue Dawid and Stone, the resulting inference about the parameter θ is liable to marginalization paradox.

This result has, *prima facie*, the trappings of non-conglomerability—a result we have come to expect given the implicit failure of countable additivity that goes with the use of improper, ignorance (prior) distributions. That is, we cannot express the marginal posterior probability density [J] as a product of a “likelihood” [K] and a prior density over θ . Thus, the posterior density for θ is not proportional to a dis-integration in θ , and Dubins has shown this equivalent to a failure of conglomerability with respect to the partition induced by θ (DUBINS, 1975).

This account is, I maintain, in error. I propose, instead, to argue that [J] does not follow from [H] (it does not follow from [I]), and that when we note why application of the usual rules for transformation of continuous random variables fails here, a simple correction avoids the paradox.

Let (x_1, x_2) be continuous, positive quantities, and consider a finitely additive credal probability function corresponding to the improper joint distribution whose density is uniform $dx_1 dx_2$. (*Note:* This density fails to determine a unique credal probability, but my point does not rest on uniqueness.) We suppose, also, that this joint density factors, according to conventional rules, into a product of a marginal and conditional density: with the intuitive results that the variables are independently, uniformly distributed. Next, transform to the pair (x_1, y) , where $y = x_2/x_1$. The calculus determines the joint density for the transformed variables: $x_1 \cdot dx_1 dy$.

But, if we apply the same conventional rules to factor this joint distribution into a product of marginal and conditional densities, the upshot is a marginal density for x_1 that is *not* uniform, but increasing in x_1 : $x_1 dx_1$. (*Note*: Thus, for this objection, we need suppose dis-integrability of the joint density in the x_1 -margin solely.)

I have argued elsewhere (SEIDENFELD, 1980) that this phenomenon is not restricted to finitely additive probability, but appears (on sets of ‘measure 0’) in the traditional theory of Kolmogorov. (The ‘Borel paradox’, uniform distribution on a sphere, illustrates the anomaly with countably additive probability.) The difficulty stems, I suggest, from the use of conventions for fixing conditional distributions when the conditioning event is the outcome of a continuous random variable, i.e. when the conditioning event has unconditional credal probability 0. (The principal villain is the rule that defines the troublesome conditional density as a limit of conditional densities each of which has a conditioning event of positive measure.) Formally, there is a simple correction to the suspect rule. In order to undo the undesirable effects of the convention (when applied to transformed joint densities), simply multiply the marginal density, as it would be obtained normally, by the Jacobian of the transformation. In this case, the Jacobian has value $J = 1/x_1$, so the corrected marginal density for x_1 , with respect to the pair (x_1, y) , is $(1/x_1)x_1 \cdot dx_1 = dx_1$, the required uniform density.

In the argument provided by Professor Stone, the pair (W, σ) is transformed to the pair (θ, σ) , where $\theta = W/\sigma$. When the marginal distribution $[J]$ for θ is obtained from the joint density $[I]$ by integrating out σ , no correction is made for the distorted marginal prior density ($d\sigma$ instead of the required $d\sigma/\sigma$) forthcoming from $[I]$. As before, the correction to $[I]$, for calculating the marginal density for σ , is to multiply by the Jacobian of the transformation ($J = 1/\sigma$); so that integrating out σ from $[I]$ (to arrive at the marginal density for θ) is in error by a factor of the Jacobian. Formally, the correct posterior marginal density for θ can be arrived at by using the conventional rules with respect to the improper joint prior density $dw \cdot d\sigma/\sigma^2$. But, as Professor Stone has shown (STONE, 1972), no marginalization paradox is generated for inverse inference about θ from this ignorance prior. Hence, I propose, no marginalization paradox is *created* by continuous transformations once the mathematics for extracting marginal densities is made coherent.

Let me summarize the results of this section. We rebuilt fiducial inference and examined its connections with Laplacean Insufficient Reason. In the

process, we saw that countable additivity was dropped and replaced by the weakened requirement of finitely additive credal probability. But finitely additive probability is compatible with non-conglomerability—hence, some of the marginalization paradoxes are explained away by noting that they are just violations of conglomerability. Finally, we pointed out that, in the presence of improper distributions, rules for transformations of continuous variables must be updated (though similar difficulties exist with proper distributions, but in much diminished scope), and revisions of these rules blocks the marginalization paradoxes created by transformations.

4. Incoherence of fiducial inference

There is one item remaining on the list of three goals for this paper: to demonstrate the failure of those hybrids, e.g. fiducial inference, structural inference, and invariance, that combine a reduction of inverse to direct inference with the Laplacean heritage—using uniform distributions to depict ignorance. The following elementary example brings us to the end of our investigation.

Suppose the problem to be determination of the volume V of a hollow cube, a problem we convert to inverse inference with data from two kinds of experiments: Relying upon our unbiased scale of known variance, $\sigma^2 = 10^{-4}$, we may fill the cube with a liquid of known density (say, unit weight/volume) and weigh this quantity of liquid: datum w_L . Also, we may cut a rigid rod of uniform density (say, unit weight/length) to the length of an edge of the cube and weigh it on the scale: datum w_R . With respect to each datum there is a smoothly invertible, canonical pivotal that may be used to arrive at a fiducial probability about V , given the respective datum:

$$v_L = (w_L - W_L); \text{ where } W_L (= V) \text{ is the weight of the liquid (unknown),}$$

$$v_R = (w_R - W_R); \text{ where } W_R (= \sqrt[3]{V}) \text{ is the weight of the rod (unknown).}$$

Using all the data, we may construct a posterior credal probability for hypotheses about V in two ways. Take either pivotal and its counterpart datum to create a fiducial probability for hypotheses about V , then use Bayes' theorem and conditionalization to add the other datum. Thus, one datum generates a probability for hypotheses about V fiducially and the other modulates this probability according to Bayes' theorem and

conditionalization through its likelihood function. However, the results of the two procedures are different posterior probabilities, a contradiction.

We can understand what has gone amiss with fiducial inference by examining the supporting Bayesian models for these alternative arguments. Remember that, for the normal distribution, the ‘ignorance’ prior that models fiducial inference with pivots v_L or v_R is uniform over possible parameter values, for the parameter appearing the pivotal. Thus, fiducial inference with v_L is modelled by an ignorance prior uniform for possible weights of the liquid, which is a uniform distribution for possible volumes V . But fiducial inference with v_R is modelled by an ignorance prior uniform for possible weights of the rod, which is a uniform distribution for possible edge lengths (of the cube), uniform over the cube-root of values of V . (Note: We may replace the fiducial argument by valid structural inference, or invariance since the example has the required group-invariance.) Just as Laplace’s principle is sensitive to alternative parameterizations, so too these hybrid programs are sensitive to which data are used to fix the parameterization that receives the privileged uniform distribution. Such sensitivity is incoherent. However, the incoherence is *not* due to the impropriety (finite additivity) that is needed with unbounded parameter spaces. We could, simply, truncate the parameter space in the problem of the hollow cube without avoiding the incoherence, though eliminating the impropriety over the bounded parameter space. (Similarly, there is no marginalization paradox in Professor Stone’s discrete example if we truncate the parameter space to the four possibilities that survive after we make an observation.) In short, I conclude that the real difficulties for programs, like Fisher’s fiducial argument, are inherited from Laplacean ideas that ignorance can be modelled by some *precise* credal probability, not from the use of finitely additive probability corresponding to impropriety.

References

- BUEHLER, R. J., and A. P. FEDDERSEN, 1963, *Note on a conditional property of Student’s t*, Annals of Mathematical Statistics, vol. 34, pp. 1098–1100
- DEMPSTER, A. P., 1963a, *On direct probabilities*, Journal of the Royal Statistical Society, Ser. B, vol. 25, pp. 100–110
- DEMPSTER, A. P., 1963b, *Further examples of inconsistencies in the fiducial argument*, Annals of Mathematical Statistics, vol. 34, pp. 884–891
- DUBINS, L. E., 1975, *Finitely additive conditional probabilities, conglomerability and disintegrations*, Annals of Probability, vol. 3, pp. 89–99

- DE FINETTI, B., 1972, *Probability, induction and statistics* (John Wiley, London), especially Chapter 5 (1949)
- FISHER, R. A., 1973, *Statistical methods and scientific inference*, (Hafner, New York), 3rd ed.
- FRASER, D. A. S., 1968, *The structure of inference* (John Wiley, New York)
- HACKING, I., 1965, *Logic of statistical inference* (Cambridge University Press, Cambridge)
- HEATH, D., and W. SUDDERTH, 1978, *On finitely additive priors, coherence, and extended admissibility*, Annals of Statistics, vol. 6, pp. 333–345
- JEFFREYS, H., 1967, *Theory of probability* (Oxford University Press, Oxford), 3rd ed.
- KOLMOGOROV, A. N., 1956, *Foundations of the theory of probability* (Chelsea, New York), 2nd English ed.
- KYBURG, H. E., 1974, *The logical foundations of statistical inference* (Reidel, Boston and Dordrecht)
- LEVI, I., 1980, *The enterprise of knowledge: An essay on knowledgee, credal probability and chance* (MIT Press, Cambridge)
- LEVI, I., *Comment on 'On some statistical paradoxe and non-conglomerability' by Bruce M. Hill* (forthcoming)
- LINDLEY, D. V., 1958, *Fiducial distributions and Bayes' Theorem*, Journal of the Royal Statistical Society, B., vol. 20–21, pp. 102–107
- LINDLEY, D. V., 1965, *Introduction to probability and statistics: from a Bayesian viewpoint* (Cambridge University Press, Cambridge), 2 volumes
- SEIDENFELD, T., 1979a, *Philosophical problems of statistical inference: learning from R. A. Fisher* (D. Reidel, Boston and Dordrecht)
- SEIDENFELD, T., 1979b, *Why I am not an objective Bayesian*, Theory and Decision, vol. 11, pp. 413–440
- SEIDENFELD, T., 1980, *Probability, continuity, and transformations of continuous random variables*, in: Essays in Semantics and Epistemology, eds. H. Leblanc, R. Gumb, and R. Stern (Heven Press, New York) (forthcoming)
- STONE, M., 1976, *Strong inconsistency from uniform priors*, Journal of the American Statistical Association, vol. 71, pp. 114–125, with discussion
- STONE, M., 1979, *Review and analysis of inconsistencies related to improper priors and finite additivity*, this volume
- STONE, M., and A. P. DAVID, 1972, *Un-Bayesian implications of improper Bayes inference in routine statistical problems*, Biometrika, vol. 59, pp. 369–375
- TUKEY, J. W., 1957, *Some examples with fiducial relevance*, Annals of Mathematical Statistics, vol. 28, pp. 687–695

REVIEW AND ANALYSIS OF SOME INCONSISTENCIES RELATED TO IMPROPER PRIORS AND FINITE ADDITIVITY

MERVYN STONE

University College, London, England

Introduction and summary

KEMPTHORNE (1976), addressing philosophers, wrote:

'Jeffreys and others coined the idea of improper prior distributions. Hacking voiced the views of many of us by asking how you could combine something that was not a probability and get a probability. It was all rather obvious. However, the use of improper priors has been pursued avidly. There has recently come to my attention a paper which I have not had time to digest by DAWID *et al.* (1973). My impression is that these workers give examples of inconsistency that result from the use of improper priors. Perhaps one should not be surprised because playing with infinities is a tricky business. There is a strong suggestion that the whole improper distribution gambit should be thrown out. It is possible that the process can be justified as a matter of approximation, but no such justification has been given, it seems.'

In the discussion of the paper by DAWID *et al.* (1973), D. V. Lindley had expressed the equally strong view: 'The paradoxes displayed here are too serious and impropriety must go'.

This paper will explicate two such paradoxes in as non-technical a manner as possible and discuss some reactions to them. Their analysis in terms of the approximability of posteriors from improper priors by those from proper priors *for the relevant class of data* is emphasized, adopting the conventional view of infinity as approximator of necessary finiteness.

The connections between impropriety and finite additivity, arising from the work of HEATH and SUDDERTH (1978), are examined to see what light can be thrown on the well-developed viewpoint of DE FINETTI (1972, 1974, 1975).

Serious impropriety in linear subatomics

The Flatland example, introduced by STONE (1976), is essentially a Bayesian version of a classical counterexample to invariant¹ statistical procedures due to Lehmann (1959, p. 24). An easily appreciated variant is provided by the following inferential game. An unemployed nuclear physicist is asked to toss a regular tetrahedral die a very large number of times. The faces of the die are labelled:

- e^+ for positron,
- e^- for electron,
- μ^+ for muon,
- μ^- for antimuon.

The outcome of each toss, that is, the label of the hidden face, is sequentially recorded, subject to the rule, readily acceded to by the physicist, that if e^+ succeeds e^- (or e^- succeeds e^+), then both symbols are annihilated, that is, the record of both outcomes is erased without trace (similarly if μ^+ follows μ^- , or if μ^- follows μ^+). The length of the record required is not specified except that it should be considerable. When the physicist says he has done enough, he is asked to do just one more toss and record the outcome in the same way as for the others. The inferential problem is simply how to assign probabilities to the outcome of the specially requested toss, given the record of the whole sequence.

A Bayesian solution requires some prior probabilities for the unknown 'parameter', θ , that is the record of labels at the stage when the physicist says he has done enough. Conditional on this parameter, the outcome of the special toss has an equiprobability distribution that may be regarded as the 'error' distribution. Together, parameter and error produce the 'observation' of the final record, x , to which Bayes theorem can be applied.

For illustration, suppose the observation x ended thus:

$$\dots e^+ \mu^+.$$

An invariant (strictly equivariant) procedure is of the sort that would analyse data on the weights of peanuts and coal deposits in the same way, apart from a 'mere' change of scale.

The likelihood function would then be 1/4 for each of the parameter values

$$\dots e^+, \dots e^+ \mu^+ e^+, \dots e^+ \mu^+ e^-, \dots e^+ \mu^+ \mu^+ \quad (1)$$

and zero for all other parameter values. If the improper uniform prior over the infinite set of parameter values of unlimited finite length is adopted, the values (1) will each get posterior probability 1/4, which will also be the posterior probability that the outcome of the special toss was any one of e^+ , e^- , μ^+ , μ^- . Yet, before examining the observation, x , we can uncontroversially assert that the probability that the special toss will give the final symbol of the observation, *without annihilation*, is 3/4. A strong inconsistency has arisen in which the marginal assignment of probability 3/4 to the event, \bar{A} , of 'non-annihilation' conflicts with its posterior probability 1/4 conditional on *any* observation.

As Seidenfeld has suggested in recent correspondence, this strong inconsistency can be regarded as an example of de Finetti's non-conglomerability concept. The basis for this is shown in Figure 1 which exposes the sleight of hand whereby a proportion of 1/4 of cases of annihilation, A , in rows is turned into a proportion of 3/4 in each member of a carefully drawn partition (just one member is illustrated). The fact that the partition happens to correspond to observations does not, I think, add much to whatever case there may be for such manipulations, when impropriety is present.

| Enumerated parameter value θ | Marginal prob. | Error outcome | | | |
|-------------------------------------|----------------|---------------|--------------|----------------|----------------|
| | | e^+ 1/4 | e^- 1/4 | μ^+ 1/4 | μ^- 1/4 |
| $\dots e^+$ | 0 | \bar{A} | A | \bar{A} | \bar{A} |
| $\dots e^+ \mu^+ e^+$ | 0 | \bar{A} | A | \bar{A} | \bar{A} |
| $\dots e^+ \mu^+ e^-$ | 0 | A | \bar{A} | \bar{A} | \bar{A} |
| $\dots e^+ \mu^+ \mu^+$ | 0 | \bar{A} | \bar{A} | \bar{A} | A |
| : | : | : | : | : | : |

Fig. 1. The probability space for the non-conglomerability associated with the improper uniform prior. The probability is uniform in any finite subset of the probability space

HILL (1979), taking the viewpoint that 'the finitely additive theory of de Finetti is the only theory without gaping holes', has commented on the Flatland example and arrived at a different conclusion, which would

presumably apply to the present discussion. He would see as a 'very weak link' the argument that, because the final toss of the tetrahedron will be recorded (without annihilation) with probability $3/4$ *conditional on each parameter sequence*, then $3/4$ is the relevant *marginal* probability, to be used for prospective evaluation. The weakness arises, because countable additivity is needed to derive the marginal probability from the infinity of conditional probabilities; it is not necessary therefore, according to Hill, to take seriously a counterexample, to the position that anything based on finite additivity is allowable, that appears to involve a countable additivity step.

Pace Hill (whose attitude appears to resemble that of a man about to suffer execution relieved to hear the executioner believes he has an *infinity* of ways of carrying it out), the unfavourable outlook for anyone placing bets on the event A , on the basis of the uniform posterior probabilities $1/4$, is clear. The apparently fair bet of 1 Eurodollar on A , with an associated profit of $1/3$ Eurodollars if A happens and the loss of the stake money if it does not, would in fact have expectation of loss of $2/3$ Eurodollars for all parameter values generated by the physicist. The posterior probabilities are therefore incoherent in the strong (and unpleasant) sense of de Finetti as interpreted by HEATH and SUDDERTH (1978).

Our next example of a Bayesian aberration stemming from the use of improper priors will not involve strong-sense incoherence. However, the results are so unattractive that they lend support to the axiomatic approaches of SHIMONY (1955) and KEMENY (1955) that adopt a weaker form of incoherence than de Finetti's.

A marginalization paradox: Treatment v. Control

If the 'difference' observations in a paired comparison of a 'treatment' and its 'control' are known to be independently and identically distributed $N(\mu, \sigma^2)$ where μ and σ^2 are unknown parameters, the probability of a 'difference' in favour of the treatment (a measure of the efficacy of the treatment) is a monotone function of $\theta = \mu/\sigma$. Thus statistical inference about θ is a problem of some practical importance. For Bayesian inference, the improper prior with element $d\mu d\sigma/\sigma$ has very impressive credentials, which now include links with finite additivity (HEATH and SUDDERTH, 1978). If, as in STONE and DAWID (1972), we *marginalize* the associated joint posterior distribution for μ and σ^2 to get the marginal posterior

distribution of θ based on a sample of differences $\underline{x} = (x_1, \dots, x_n)$, the posterior density is

$$\pi(\theta|\underline{x}) \propto \exp(-\frac{1}{2}n\theta^2) \int_0^\infty \omega^{n-2} \exp(-\frac{1}{2}\omega^2 + r\theta\omega) d\omega, \quad (2)$$

where $r = (\sum x_i)/(\sum x_i^2)^{\frac{1}{2}}$. The *marginalization paradox* in this result is simply that

- (a) $\pi(\theta|\underline{x})$ is a function of n and r only, implying that we need only be informed of the values of n and r to construct the posterior distribution of the parameter of interest,
- (b) a Bayesian so informed could never agree that (2) represents Bayesian inference.

The reason that (2) must be certified ‘un-Bayesian’ is that r has a distribution with density

$$f(r|\mu, \sigma) = f(r|\theta) \propto \left(1 - \frac{r^2}{n}\right)^{\frac{1}{2}(n-3)} \int_0^\infty \omega^{n-1} \exp(-\frac{1}{2}\omega^2 + r\theta\omega) d\omega \quad (3)$$

that depends only on θ and that cannot be combined with any prior for θ to match (2); the inconsistent powers of ω in the two integrands prevent that happy reconciliation from taking place.

ROBINSON (1978) has even proposed a measure of the un-Bayesianity of induction based on (2). In this case the measure takes the form

$$S = \sup_{r_1, r_2, \theta_1, \theta_2} \left\{ \frac{\pi(\theta_1|r_1) f(r_2|\theta_1) f(r_1|\theta_2) \pi(\theta_2|r_2)}{f(r_1|\theta_1) \pi(\theta_1|r_2) \pi(\theta_2|r_1) f(r_2|\theta_2)} \right\}.$$

I have not evaluated S for any n ; however, for $n = 2$, $r_1 = 0$, $r_2 = 1$, $\theta_1 = 0$, $\theta_2 = 1$, the ratio inside the brackets is 1.62, so that S for $n = 2$ is at least this large. A value of S not equal to 1 means that there is at least one pair of values of θ for which the ratio of posterior densities is not proportional (with respect to different data sets) to the likelihood ratio, as the Bayes theorem requires. In fact, in the present example, this breakdown of Bayesianity occurs for all pairs of values.

BERNARDO (1979) does not deny the paradoxical nature of the marginalization paradox, but devises a procedure that appears to avoid it by specifying a different ‘reference’ ignorance prior for each parameter of interest.

The implications for finite additivity of the marginalization paradox can be established through the recent work of HEATH and SUDDERTH (1978). The conditions of Section 4 and Corollary 2 of their paper hold and imply that the joint posterior for μ and σ^2 and *a fortiori* the marginal posterior for θ are *weakly coherent*, that is, not strongly incoherent in the sense of de Finetti. So it is not possible simultaneously to maintain that weak coherence is acceptable and to reject its consequences in (2). However, I conjecture that, in addition to being weakly coherent, the inductive probability distributions (2) may be shown to be *weakly incoherent* in the sense previously indicated; more precisely, in the sense obtained by relaxing the requirement of *uniform* positivity (but not the positivity) of expectation of payoff in HEATH and SUDDERTH's (1978) inequality (2.1). (This uniform positivity corresponds to the expectation of loss being uniformly 2/3 Eurodollars in our first example.)

Just as his firm support of finite additivity would lead HILL (*loc. cit.*) to reject the implications I have drawn in the first of the present examples, so a conviction that improper priors cannot be seriously misleading (at least for statistical inference) has led JAYNES (1979) to reject any adverse implications of the marginalization paradox. In reply to the discussion of his paper by DAWID *et al.*, JAYNES (*loc. cit.*) introduces the useful concept of an improper prior as a limit, in some sense, of a sequence of proper priors in order to argue, however, that (2) is acceptable. In view of the rather technical character of our second example, it is preferable, as well as adequate, to consider the logic and consequences of Jaynes's mode of argument by applying it to the simpler first example. Instead of an improper uniform prior over the whole set Θ of parameter values, θ , we assign the proper prior probabilities

$$\pi_p(\theta) = \exp[-L(\theta)^2/p] / \sum_{\theta \in \Theta} \exp[-L(\theta)^2/p], \quad (4)$$

where $L(\theta)$ denotes the length of θ and p is a positive integer. It is irrefutable that, for *fixed* observation, the limit as $p \rightarrow \infty$ of the posterior probabilities obtained by using π_p is 1/4, their value for the uniform improper prior. This, no more and no less, is Jaynes's mode of argument for the acceptability of posteriors derived from improper priors. (A different example, with further comment, is to be found in STONE (1976).)

Despite our rejection of Jaynes's use of asymptotics in justification of improper priors, it is possible to agree with the final sentences of Jaynes's reply: 'For a practical Statistician, the notion of an improper prior is

Often a natural and useful idealization—just as the notion of a perfect triangle is for a surveyor. For both, it is part of their professional competence to understand clearly under *which* conditions the idealization is appropriate'. We now explore some of the ways of achieving such understanding.

Infinity—limits and limitations

The ‘double or quits’ betting strategy is, for very sound reasons, less used as a guide for action than as generator of the well-known St. Petersburg paradox: ‘For fair-coin betting, staking 2^{n-1} Eurodollars on the n th toss will yield a profit of 1 Eurodollar when a Head comes up, which it will do with probability one.’

The paradox reveals how an event of probability zero (the event of no Heads in an infinity of tosses) can vitiate the asymptotic analysis of a problem in decision theory. It also shows the inadequacy of weak convergence for some applications (of which more later). We should therefore not be surprised to run into problems when an event *of probability one* is neglected, which, surprising as it may seem, may be the consequence of some asymptotic justifications in this area.

To start with, consider again the problem of inference in the Linear Subatomics example. The issue is more clearly brought out if, instead of using proper prior distributions of the type (4), we take the sequence of proper prior distributions $\{\pi_p\}$ in which π_p assigns equal prior probability to all the parameter values of length less than or equal to p . As $p \rightarrow \infty$ for *fixed* observation, we have the same apparent justification of the strongly incoherent posteriors as was provided by (4). Given that we are unhappy about such justification, we may express our puzzlement in the question, “Where has the Bayesianity gone?”.

The paradox is resolved by directing attention to the form of the posterior distributions corresponding to observations of length p and $p+1$. For each of the latter, the posterior probability is concentrated on the single parameter value that is consistent with the observation and that has positive prior probability. The crucial point now is that the proportion of potential observations up to length $p+1$ that are of length p or $p+1$ does not tend to zero as $p \rightarrow \infty$: in other words, we cannot neglect asymptotically, as an ‘edge effect’, those observations that give radically different posteriors from the equi-probability posterior that is generated as the

limit for each fixed observation. In fact, the marginal probability of an observation in the anomalous category (length p or $p+1$) is

$$\frac{2}{3(1-1/3^p)}. \quad (5)$$

There is nothing un-Bayesian in assigning the equiprobability posterior distribution to every observation of length up to $p-1$, provided we also do something such as employing the above degenerate posterior distributions for observations of length p or $p+1$. *That* is where the Bayesianity went! It should be noted that the question at issue here is not the reasonableness of the prior distributions involved but the logic of their inferential implications.

In the example just considered, we see that the improper prior in effect neglects a portion of the observation space with asymptotic marginal probability $2/3$, the limit of (5) as $p \rightarrow \infty$.

For our second example, we return to the case of Treatment v. Control and find that an even larger asymptotic portion of observation space is neglected, although the results of the neglect are, we have conjectured, only weakly, as opposed to strongly, incoherent. The analysis for this example is necessarily somewhat more technical and for the details, which we will here merely summarize, we refer the reader to STONE and DAWID (1972). As there demonstrated, a sequence $\{\pi_p\}$ of proper prior distributions for μ and σ^2 could be exhibited with the following properties:

- I. The joint posterior distribution of μ and σ^2 for *fixed* data $x = (x_1, \dots, x_n)$ using the prior π_p tends, as $p \rightarrow \infty$, to the formal joint posterior given by the improper prior $d\mu d\sigma/\sigma$.
- II. The joint posterior of the pivotal quantities

$$t = \sqrt{n}(\bar{x} - \mu)/s, \quad \chi^2 = (n-1)s^2/\sigma^2,$$

where $s^2 = \sum(x_i - \bar{x})^2/(n-1)$, tends, as $p \rightarrow \infty$, in marginal probability, with respect to the total variation metric, to their formal joint posterior using $d\mu d\sigma/\sigma$ (which happens to be the same for all x).

- III. The statistic $r^2 = (\sum x_i)^2 / \sum x_i^2$, whose value was influential in the marginalization paradox, tends to n in marginal probability as $p \rightarrow \infty$, which is equivalent to $|\bar{x}/s|$ tending to ∞ .

In fact, III is a necessary condition for II. Property I could be the basis of a Jaynesian justification of the unattractive (2). However, III shows that, at least for any sequence of proper priors having the statistically acceptable property II, the marginal probability of any region of observation space with bounded $|\bar{x}/s|$ tends to zero as $p \rightarrow \infty$, that is, such justification would effectively be asymptotically neglecting an event of probability one!

AKAIKE (1978) has developed quite a different approach to the analysis of the marginalization paradox by means of a sequence of proper priors. Akaike believes he has exposed 'the fallacy of interpreting an improper prior distribution as a limit of some proper prior distributions' and that he has 'shown that the improper prior distribution can more adequately be described as the limit of some *data adaptive* [my italics] proper prior distributions'. He maintains that 'a prior distribution without data adaptability cannot escape the risk of producing extremely poor inference due to a gross mis-specification of the prior'. Although many statisticians might concur with this opinion, a line of defence of the marginalization paradox impropriety, that essentially argues that there is no significant paradox because strict Bayesianity is risky, is not likely to commend itself to rigorous Bayesians.

For problems involving well-behaved (amenable) groups (e.g. 'Treatment v. Control' but not 'Linear Subatomics') a mathematically interesting alternative to direct consideration of the posterior for a sequence of proper priors stems from the work of BONDAR (1977b) and HEATH and SUDDERTH (1978). To allow the concept of 'convergence' to emerge, if only superficially as we will see, it is necessary to replace the sequence of proper priors by a *net* of them. (The 'Appendix on general topology' in ASH (1972) is particularly concise and useful.) The net is constructed to converge to a finitely additive probability distribution π^* which has completely specified properties and which may be regarded as corresponding to those of a uniform (Haar) prior on the group of parameter values. (An arbitrariness of notation accounts for the fact that the same Haar prior is *right* invariant in STONE (1970) and *left* invariant in HEATH and SUDDERTH (1978).) Unfortunately, as Heath and Sudderth acknowledge, there are infinitely many such limits π^* , each of which has the status of an invariant mean (GREENLEAF, 1969). Their non-uniqueness may be enough in itself to raise some doubts about the robustness of this application of finitely additive probability distributions.

In order to illustrate the alternative approach and to gain some insight

into the difficulties entailed, at least in relation to the construction of π^* if not its principal application, it is happily sufficient to treat the problem of choosing an integer at random from the extended parameter set

$$\Theta = \{\theta\} = \{\dots, -2, -1, 0, 1, 2, \dots\} \quad (6)$$

which is a problem whose realism may be regarded as a touchstone of the finite additivity viewpoint. To this end, let π_p denote the proper prior that attaches probability $1/p$ to the θ -values $1, \dots, p$ and zero probability to all other values of θ . Let π denote a general finitely additive probability distribution on Θ . Endow the space $\Pi = \{\pi\}$ with the (weak*) topology T in which the open sets have for base the cylinder sets of the general form

$$\{\pi \mid |\pi(\Theta_i) - \pi_0(\Theta_i)| < \varepsilon, i = 1, \dots, k\},$$

where $\varepsilon > 0$, $\pi_0 \in \Pi$ and $\Theta_i \subset \Theta$, $i = 1, \dots, k$ with k finite. Then, with respect to the topology T , Π is compact but not first countable. The compactness implies that the sequence $\{\pi_p\}$ has at least one accumulation point in Π . Selecting one of these as π^* , we have that, given $\varepsilon > 0$ and subsets $\Theta_1, \dots, \Theta_k$ of Θ , there are arbitrarily large values of p with

$$|\pi_p(\Theta_i) - \pi^*(\Theta_i)| < \varepsilon, \quad i = 1, \dots, k. \quad (7)$$

Can one be entirely happy with (7) as an asymptotic justification of π^* ? I suggest not. Roughly, (7) tells us that, given an interest in $\Theta_1, \dots, \Theta_k$, there are an infinity of p values with corresponding comprehensible proper prior distributions for which the prior probabilities of $\Theta_1, \dots, \Theta_k$ are as close as we like to those assigned by π^* . It does not tell us what is happening to $|\pi_p(\Theta_{k+1}) - \pi^*(\Theta_{k+1})|$ for these values of p when Θ_{k+1} is not equal to any of $\Theta_1, \dots, \Theta_k$. In fact, we can construct Θ_{k+1} such that $\pi^*(\Theta_{k+1})$ is a uniformly poor approximation of $\pi_p(\Theta_{k+1})$ for an infinity of the p -values that served satisfactorily for $\Theta_1, \dots, \Theta_k$.

To see that the introduction of nets in place of sequences (as in HEATH and SUDDERTH, 1978) does not resolve this difficulty, even though it allows the word ‘convergence’ to replace the word ‘accumulation’, we let D denote the collection of all pairs (p, V) where V is a neighbourhood of π^* such that $\pi_p \in V$. Then D is made a *directed* set by defining

$$(p_1, V_1) \leq (p_2, V_2) \Leftrightarrow p_1 \leq p_2 \text{ and } V_2 \subset V_1.$$

For $d = (p, V)$, define $\pi_{p_d} = \pi_p$. Then it can easily be shown that $\{\pi_{p_d}, d \in D\}$ is a subnet of $\{\pi_p\}$ that converges to π^* . This convergence means that, given any neighbourhood U of π^* , there is a (p_0, V_0) such

that, for all values of $p \geq p_0$ for which there is $(p, V) \in D$ with $V \subset V_0$, we have $\pi_p \in U$. In particular, taking U to be $\{\pi \mid |\pi(\Theta_i) - \pi^*(\Theta_i)| < \varepsilon, i = 1, \dots, k\}$, it means *no more* than the accumulation behaviour expressed in (7). (Take p_0 to be the smallest value of p satisfying (7) and $V_0 = U$.) This appears to be a case of change of mathematical terminology without change of substance.

The results of BONDAR (1977b) and HEATH and SUDDERTH (1978) show that any such problems² over the justification of π^* as a representation of uniform probability over Θ do not interfere with the principal application of π^* , which is to statistical inference about θ when θ is made the location parameter of an appropriately defined statistical model. For our simple example (6), suppose that a real observation x has the two point discrete distribution

$$P(x = \theta - 1|\theta) = P(x = \theta + 1|\theta) = \frac{1}{2}. \quad (8)$$

For such an observation, the uniform improper prior distribution over Θ yields the formal posterior

$$\pi(\theta = x - 1|x) = \pi(\theta = x + 1|x) = \frac{1}{2}. \quad (9)$$

The application of the theory (as in HEATH and SUDDERTH, 1978) associated with π^* then assures us that the posteriors (9) are weakly coherent (that is, not strongly incoherent in the de Finetti sense). The key steps in the general theory are:

- (a) the definition of a marginal finitely additive probability distribution m^* for x , as the expectation with regard to π^* of the general analogue of the conditional probability (8);
- (b) a tightening of the specification of π^* by additional conditions on the sequence $\{\pi_p\}$ of which π^* is an accumulation point. (In the current example, the postulated $\{\pi_p\}$ would satisfy the conditions but $\{\pi_p\}$ in which each π_p is, for example, concentrated on the integers not divisible by 3 would not);
- (c) the demonstration of what might be called ‘expectation equivalence’, that is, the expectation with respect to π^* of the *forward* conditional expectation of any bounded function $\Phi(x, \theta)$ given θ equals the expectation with respect to m^* of the *inverse* conditional expectation of $\Phi(x, \theta)$ given x .

² Note also that the π^* we have constructed will give zero probability to any set of non-positive values of θ .

HEATH and SUDDERTH (1978) show that (c) is equivalent to weak coherence. Example 6 of BUEHLER (1976) is equivalent to exhibiting a bounded $\Phi(x, \theta)$ whose forward expectation with respect to (8) is negative for all θ (but not uniformly negative) but whose inverse expectation with respect to (9) is zero for all x . This example also reminds us that weak incoherence is by no means the prerogative of the marginalization paradox.

The role of π^* and m^* is to deny the importance of such examples by keeping sights on infinity, as it were. In this respect, they relate precisely to the type of asymptotic justification illustrated in II and III above, in connection with the Treatment v. Control paradox. This application of finite additivity serves to justify uniform impropriety in certain contexts where it is believed that π^* or something close to it has an operational role in the generation of θ and the x associated with it. If on the other hand, there is not a wholehearted commitment to the infinity implications for θ and x , the marginalization paradox shows the danger of allowing the Bayesian baby to be lost on the infinitely spreading bath water and of being left with an unattractive residue.

Acknowledgements

I wish to acknowledge useful discussion with D. G. Larman, University College London, and valuable assistance from J. V. Bondar for the topological content of this paper.

References

- AKAIKE, H., 1978, *The use of improper prior distributions as limits of data dependent proper prior distributions*, Res. Memo. 137 (Institute of Statistical Mathematics, Tokyo)
- ASH, R. B., 1972, *Real analysis and probability* (Academic Press, New York and London)
- BERNARDO, J. M., 1979, *Reference posterior distributions for Bayesian inference*, With discussion, Journal of the Royal Statistical Society, B vol. 41, pp. 113–147
- BONDAR, J. V., 1977a, *Some principles of statistical inference and three examples in which they fail*, Personal manuscript
- BONDAR, J. V., 1977b, *The mean of Neyman-Pearson risk equals the mean of invariant Bayes risk*, Personal manuscript
- BONDAR, J. V., 1977c, *A conditional confidence principle*, Annals of Statistics, vol. 5, pp. 881–891
- BONDAR, J. V., and P. MILNES, 1976, *A survey of Hunt-Stein and related conditions on groups*, Submitted to Annals of Statistics
- BUEHLER, R. J., 1976, *Coherent preferences*, Annals of Statistics, vol. 6, pp. 1051–1064

- DAWID, A. P., and M. STONE, 1972, *Expectation consistency of inverse probability distributions*, Biometrika, vol. 59, pp. 486–489
- DAWID, A. P., and M. STONE, 1972, *Expectation consistency and generalized Bayes inference*, Annals of Statistics, vol. 1, pp. 478–485
- DAWID, A. P., M. STONE, and J. V. ZIDEK, 1973, *Marginalization paradoxes in Bayesian and structural inference*, With discussion. Journal of the Royal Statistical Society, B, vol. 35, pp. 189–233
- DAWID, A. P., M. STONE and J. V. ZIDEK, 1979, *Comments on Jaynes' paper "Marginalization and prior probabilities"*, in: Bayesian analysis in econometrics and statistics; Essays in honor of Harold Jeffreys, ed. A. Zellner (North Holland, 1980), pp. 79–82
- DE FINETTI, B., 1972, *Probability, induction and statistics* (John Wiley & Sons)
- DE FINETTI, B., 1974, *Theory of probability*, vol. I (John Wiley & Sons)
- DE FINETTI, B., 1975, *Theory of probability*, vol. II (John Wiley & Sons)
- DUBINS, L. E., 1975, *Finitely additive conditional probabilities, conglomerability and disintegration*, Annals of Probability, vol. 3, pp. 89–99
- GREENLEAF, F. P., 1969, *Invariant means on topological groups and their applications* (Van Nostrand, Reinhold Co.)
- HEATH, D., and W. SUDDERTH, 1978, *On finitely additive priors, coherence and extended admissibility*, Annals of Statistics, vol. 6, pp. 333–345
- HILL, B. M., 1979, *On some statistical paradoxes and non-conglomerability*, Invited paper at the International Meeting on Bayesian Statistics (Valencia)
- JAYNES, E. T., 1979, *Marginalization and prior probabilities*, in: Bayesian analysis in econometrics and statistics; Essays in honor of Harold Jeffreys, ed. A. Zellner (North Holland 1980), pp. 43–78
- KEMENY, J. G., 1955, *Fair bets and inductive probabilities*, The Journal of Symbolic Logic, vol. 20, pp. 263–273
- KEMPTHORNE, O., 1976, *Statistics and the philosophers*, in: Foundations of probability theory, statistical inference and statistical theories of science, eds. W. L. Harper and C. A. Hooker, vol. II (D. Reidel Publishing Co., Dordrecht–Holland), pp. 273–314
- LEHMANN, E. L., 1959, *Testing statistical hypotheses* (John Wiley & Sons, New York and London)
- ROBINSON, G. K., 1978, *On the necessity of Bayesian inference and the construction of measures of nearness to Bayesian form*, Biometrika, vol. 65, pp. 49–52
- SHIMONY, A., 1955, *Coherence and the axioms of confirmation*, The Journal of Symbolic Logic, vol. 20, pp. 1–28
- STONE, M., 1963, *The posterior t distribution*, Annals of Mathematical Statistics, vol. 34, pp. 568–573
- STONE, M., 1964, *Comments on a posterior distribution of Geisser and Cornfield*, Journal of the Royal Statistical Society, B, vol. 26, pp. 274–276
- STONE, M., 1965, *Right Haar measure for convergence in probability to quasi posterior distributions*, Annals of Mathematical Statistics, vol. 36, pp. 440–453
- STONE, M., 1970, *Necessary and sufficient condition for convergence in probability to invariant posterior distributions*, Annals of Mathematical Statistics, vol. 41, pp. 1349–1353
- STONE, M., 1976, *Strong inconsistency from uniform priors*, With discussion. Journal of the American Statistical Association, vol. 71, pp. 114–125

- STONE, M., and A. P. DAWID, 1972, *Un-Bayesian implications of improper Bayes inference in routine statistical problems*, Biometrika, vol. 59, pp. 369–375
- STONE, M., and B. G. F. SPRINGER, 1963, *A paradox involving quasi prior distributions*, Biometrika, vol. 52, pp. 623–627
- STONE, M., and R. VON RANDOW, 1968, *Statistically inspired conditions on the group structure of invariant experiments and their relationships with other conditions on locally compact topological groups*, Z. Wahrscheinlichkeitstheorie und verw. Geb., 10, pp. 70–80

ON THE PROBLEM OF IRREVERSIBILITY IN THEORETICAL PHYSICS

A. P. GRECOS and I. PRIGOGINE *

Faculté des Sciences, Université Libre de Bruxelles, Brussels, Belgium

1. Introduction

In the theoretical description of nature we often appeal, explicitly or implicitly, to the second law of thermodynamics. This is a formal statement about the experimental fact that spontaneous processes occurring in physical systems are irreversible.

The second law deals with the dissipation of energy and is logically independent of the law of conservation of energy. This was clearly understood in the middle of the nineteenth century and R. CLAUSIUS (1865) introduced the term *entropy* as a “measure of transformability”. He also stressed the generality of the second law in a famous sentence: “The entropy of the world tends toward a maximum”.¹

Thus the second law postulates that isolated (macroscopic) systems eventually reach a state of equilibrium because of irreversible processes. Moreover, this tendency is characterized by the existence of the entropy S , that is a non-decreasing function of the variables defining the state. For non-isolated systems the entropy change dS is the sum of two terms: the flow of entropy $d_e S$ across the boundaries of the system and the entropy production $d_i S$ due to internal processes (GLANSDORFF and PRIGOGINE, 1971). In this case the second law states that the entropy production is non-negative. The basic distinction here is between reversible and irreversible processes, the latter implying a specific direction in the temporal evolution of physical systems.

* Also at the Center for Statistical Mechanics and Thermodynamics, The University of Texas at Austin, Austin, Texas 78712, USA.

“Die Energie der Welt ist constant, Die Entropie der Welt strebt einem Maximum zu”.

Today the role of thermodynamical considerations in studying physical systems far from equilibrium is quite important. The introduction of the notion of "dissipative structures" has led to a unified description of coherent processes which are of particular interest in physics, chemistry and biology. Such structures are characterized by an increase of entropy production, which is minimum for stationary states near equilibrium, and are highly organized. Therefore, under certain conditions, irreversible processes may be viewed as a source of order (PRIGOGINE, 1978, and references cited there).

To formulate thermodynamics it is not necessary to refer to a microscopic theory for the structure of matter. However, it is evident that certain statements, and in particular the second law, are rather qualitative. In fact, there is no general prescription for constructing the entropy function. Thus it is natural to search for a more fundamental understanding of the nature of irreversible processes based on a dynamical theory.

By a *dynamical theory* we mean a theory where the properties of a macroscopic system are deduced from a mechanical model involving a "large" number of degrees of freedom. This is essentially the principal aim of statistical mechanics. An example of such a model is a fluid composed of $N (10^{23})$ interacting point particles, the motion of which is governed by the laws of classical or quantum mechanics. Here we meet the main conceptual difficulty. All known fundamental interactions are such that the motion is reversible with respect to time. Therefore the question arises whether, and in which sense, irreversible processes are compatible with the laws of mechanics. It is a problem that is far from being settled and it is hardly necessary to insist on its importance for the foundations of theoretical physics.

There are several aspects of the problem of irreversibility but in this paper we consider mainly those connected with the concept of entropy and the microscopic interpretation of the second law. Indeed, one of the important questions that we need to solve in constructing a dynamical theory of dissipative phenomena is the definition of a functional of the state that increases monotonically in time. Recent investigations on this matter have proven that under certain conditions such functionals exist. Our purpose is to present briefly the essential ideas that have been developed and which lead to the conclusion that irreversibility is an objective property of sufficiently complex dynamical systems. (For a more detailed review see PRIGOGINE and GRECOS, 1979.)

2. Kinetic theory

In the development of a microscopic theory of irreversible processes, the work of Maxwell and Boltzmann is of basic importance. For the first time a coherent probabilistic description of a dilute classical gas is formulated.² It is assumed that the system consists of a large number of identical particles of mass m interacting with short-range forces. As the density is taken to be small, the mean free path of the particles is much larger than the range of the forces. Macroscopic quantities are interpreted as averages of mechanical ones; in particular, heat is the average kinetic energy of the “random” motion of the molecules.

Instead of considering the mechanical state of the system, that is the set of the positions q_r and momenta p_r of the particles, a coarser description is used by introducing the one particle distribution $f(p, q; t)$. This (smooth) function represents an average density of particles in the six-dimensional space of positions and momenta. Using probabilistic and dynamical considerations, Boltzmann was able to derive for the evolution of $f(p, q; t)$ a closed equation (BOLTZMANN, 1872, 1896, 1898), namely

$$\frac{\partial f}{\partial t} + \frac{p}{m} \cdot \frac{\partial f}{\partial q} = B(f, f). \quad (1)$$

According to this theory, the change in time of the one-particle distribution is due to the free motion of the particles, given by the second term in the left-hand side of equation (1) and their (binary) collisions, denoted by $B(f, f)$ in the right-hand side of the equation. We do not need here the explicit form of the collision term of the Boltzmann equation (see GRAD, 1958). The fundamental hypothesis in the derivation is that certain spatial and velocity correlations may be neglected.

It is a consequence of equation (1) that a non-increasing functional of the state can be defined. Indeed, the famous *H-theorem* asserts that

$$\mathcal{H}(t) = \int dp dq f \log f, \quad d\mathcal{H}/dt \leq 0. \quad (2)$$

The equality is valid for the Maxwellian distribution of a gas at equilibrium. Thus $S = -k\mathcal{H}$, where k is a constant (the *Boltzmann constant*), provides a suitable “mechanical” definition of non-equilibrium entropy for a dilute gas.

* The historical process that led to the statistical-mechanical conception of thermal phenomena is extremely interesting. An excellent monograph by S. G. BRUSH (1976) has been published recently dealing with this subject.

As it is well known, the derivation of the H -theorem was a source of debates and controversies as to the meaning and the validity of the Boltzmann equation. The objections raised by Loschmidt and Zermelo (and Poincaré), based on the reversibility and recurrence of the motion of mechanical systems (LOSCHMIDT, 1876, 1877; ZERMELO, 1896a, 1896b; POINCARÉ, 1893), illustrate the fact that Boltzmann's theory cannot be considered as a direct consequence of the underlying dynamics of the model.

Of course, today the importance of Boltzmann's ideas cannot be denied. They are fundamental in the investigations of Planck and Einstein on irreversible processes involving radiation (PLANCK, 1900; see also SCHÖPF, 1978; EINSTEIN, 1917), investigations that changed profoundly our picture of the physical world. Nevertheless, the necessity of introducing a probabilistic description of dynamical systems as well as the conditions that eventually permit to establish general H -theorems require further investigation.

3. Dynamical systems

The motion of a classical system of s degrees of freedom is often conveniently represented as a trajectory in a $2s$ -dimensional space, the so-called *phase space*. For example, in the case of a system of N point particles, the *mechanical state* is the set of coordinates and momenta $x_i = \{q_1(t), p_1(t); \dots; q_N(t), p_N(t)\}$, and it may be represented as a point in a $6N$ -dimensional space. As a function of time, x_i traces a curve in the phase space determined by Hamilton's equations

$$\frac{dq_r}{dt} = \frac{\partial H}{\partial p_r}, \quad \frac{dp_r}{dt} = -\frac{\partial H}{\partial q_r} \quad (r = 1, 2, \dots, N), \quad (3)$$

where $H(p, q)$ is the energy of the system.³ Because of the conservation of energy the trajectory lies on the hypersurface $H(p, q) = \text{constant}$

In several cases the Hamiltonian is invariant when the momenta are inverted, $H(p, q) = H(-p, q)$, and thus to every solution of equation (3) there corresponds another one $\{q_r(-t), -p_r(-t)\}$. This is a manifestation

³ For a system of particles, H is the sum of the kinetic energy $\frac{1}{2m} \sum_r p_r^2$ and of the potential energy $\frac{1}{2} \sum_{r \neq s} U(q_r - q_s)$ due to interparticle interactions.

of the dynamical reversibility of the motion. Here it should be noted that Hamiltonian particle models are not the only systems of interest. An-harmonically coupled oscillators described by action-angle variables, or non-holonomic systems such as a set of hard spheres in a box, are examples of models frequently used in statistical mechanics. In any case, dynamical reversibility is a general property of isolated mechanical systems.

A statistical description is introduced by considering instead of a point in phase space, an *ensemble* (GIBBS, 1902) defined by a probability density ϱ . The evolution of this distribution is governed by the Liouville equation

$$i \frac{\partial \varrho}{\partial t} = i \{H, \varrho\} \equiv L\varrho \Leftrightarrow \varrho_t = \exp(-iLt)\varrho_{t=0} \quad (4)$$

where L is a differential operator defined by the Poisson bracket $i\{H, \cdot\}$. Probabilistic considerations do not lead directly to irreversibility. The evolution predicted by the Liouville equation is reversible, a fact expressed mathematically by the unitarity of $\exp(-iLt)$.

One may ask at this point what is the meaning of a probabilistic description of a deterministic system. It is often assumed that such a description is necessary because of our practical inability to compute the details of the mechanical state. This argument is to some extent correct but it does not explain the reason that permits statistical mechanics to make valid predictions.

Any measurement determines the value of a physical quantity with finite accuracy. Thus the notion of a state in classical mechanics as a point in the phase space is an idealization involving a limiting procedure (a measurement of absolute accuracy). However, such an idealization cannot be physically meaningful unless the motion is stable.

During the past few years, several important mathematical investigations on the properties of dynamical systems have appeared. It has been shown that the motion of classical systems is, in general, structurally unstable. Two cases should be mentioned here. The first one is the work of Sinai, showing that systems with elastic collisions, such as hard spheres in a box, are highly unstable and have strongly stochastic properties (SINAI, 1967). On the other hand, for Hamiltonian systems the situation is quite different. It was thought for some time that any interaction would destroy all (isolating) integrals of motion of an integrable system (except the energy). That this is not so, it was established some years ago by Kolmogorov, Arnold and Moser (K.A.M.-theory) (ARNOLD and AVEZ, 1968; MOSER,

1973). However, through every region of the phase space, no matter how small, qualitatively different trajectories pass. For instance, if through some point passes a periodic trajectory, it is possible to find in any neighbourhood around it, points through which pass erratic trajectories that fill regions of the phase space.⁴

The preceding remarks are intended to show that the significance of trajectories for systems even with relatively few degrees of freedom is, in general, limited because of their instability. As a consequence, the idea that classical mechanics implies an absolute determinism cannot be maintained. A statistical description becomes essential in formulating a mathematical theory of evolution.

In quantum mechanics a probabilistic reasoning is introduced *ab initio* by Born's interpretation of the wave function. Moreover, the "mechanical" state now is a vector ψ in some, usually infinite-dimensional, space. Nevertheless, with respect to the problem of irreversibility the situation is quite similar to that in classical mechanics. Schrödinger's equation for the evolution of the state ψ is time-reversible. Furthermore, another statistical elements appears by considering *mixed states*, described by density matrices, that cannot be reduced to *pure states*, described by wave vectors. It is well known (VON NEUMANN, 1955) that a density matrix ϱ evolves according to the von Neumann equation, which has the same form as equation (4) with L defined by the commutator of the Hamiltonian ($L\varrho = [H, \varrho] \equiv H\varrho - \varrho H$). As in the classical case, the von Neumann equation is time-reversible and the evolution of ϱ is determined by a one-parameter group of unitary transformations.

The use of mixed states in quantum mechanics is necessary in any attempt to develop a theory of macroscopic phenomena. Therefore, here also the question arises as to the justification of introducing probabilities through mixtures, a question that is independent of the interpretation of the wave function. It may be conjectured that the fundamental reason for such statistical considerations lies in some kind of instability of the motion. Although it is clear that such a situation cannot arise unless the spectrum of the Hamiltonian is continuous, there are no definitive results concerning unstable motions of quantum systems.

⁴ These results have revived the interest in classical mechanics and ergodic theory (see e. g. the review articles in WIGHTMAN, 1971; LEBOWITZ and PENROSE, 1973; FORD, 1973; WHITEMAN, 1973).

4. Existence of entropy functionals

Several investigations in non-equilibrium statistical mechanics deal with the conditions that permit to establish an irreversible equation for a "part" of the statistical state of a dynamical system (cf. GRECOS, 1978, PRIGOGINE and GRECOS, 1979). Thus, for example, one considers the possibility of deriving an equation of the Boltzmann type for the one-particle distribution, or an equation of the Pauli type for the diagonal elements of the density matrix. It is impossible to discuss here these questions but let us summarize briefly certain conclusions.

When the concept of the "collision term" is appropriately generalized, a kinetic description can be formulated that is exact but incomplete. This means that it describes the time development of a part of the state, which, however, cannot be completely separated from the evolution of the system as a whole. In certain cases (e.g. for a dilute gas of density c), one may show that specific equations are asymptotically ($t \rightarrow \infty$) and non-uniformly ($c \rightarrow 0$, ct : finite) valid. It is often necessary, e.g. in order to have finite transport properties, to consider the limit of an infinite system. Of course, such a limit is an idealization that permits us to make some statements precise, but one should keep in mind that the notion of a finite *isolated* system is an idealization as well. For infinite systems it is possible to impose certain restrictions on the classes of initial conditions that lead to an irreversible behaviour. In particular, statistical correlations should be of short range, as those arising from molecular interactions.

Perhaps the most essential condition that arises in any coherent discussion of a dynamical theory of irreversibility is the necessity of a Liouville, or von Neumann, operator L with a continuous spectrum. This may seem just a mathematical statement without much physical content. However, at least for classical systems, it is intimately related to the instability of the phase trajectories that we mentioned in the previous section. Finite quantum systems have a von Neumann operator with discrete spectrum and to discuss irreversible processes the limit of an infinite system is necessary. On the contrary, finite classical systems may have a Liouville operator with a continuous spectrum, the case of harmonic oscillators being an exception rather than the rule.

Let us now discuss the problem of a microscopic definition of entropy. As we have already noticed, such a definition implies the existence of functionals of the microscopic state that monotonically increase in time. It is customary in statistical mechanics to consider functionals that de-

crease in time ("H-functionals") but obviously this is simply a question of the choice of sign.

If $\Omega(\varrho)$ is an H-functional, we demand that

$$\Omega(\varrho) \geq 0, \quad d\Omega/dt \leq 0. \quad (5)$$

The unitarity of the evolution implies that the norm of ϱ ($= \int dp dq \varrho^2$ or $\text{Tr} \varrho^2$) is a constant. Similarly the functional $\mathcal{H}_G = \int_r dp dq \varrho \log \varrho$ (or $\text{Tr} \varrho \log \varrho$), proposed by Gibbs, is also a constant and cannot be used to define the entropy for systems not in (statistical) equilibrium.

It is an easy matter to prove that the inequality in (5) cannot be satisfied by linear functionals, an observation that is essentially due to Poincaré. However, under certain conditions quadratic functionals

$$\Omega = \int_I dp dq \varrho M \varrho \quad (\text{or } \text{Tr} \varrho M \varrho) \quad (6)$$

exist, for which the inequality holds.

Taking into account the Liouville (or von Neumann) equation, it follows that the operator M must be such that

$$i(ML - LM) = D \leq 0. \quad (7)$$

The condition that D is non-positive is quite restrictive. Indeed if the spectrum of L is discrete D vanishes identically. On the other hand, if the spectrum of L is continuous,⁵ non-trivial operators exist. As a matter of fact they can be constructed as a (positive) decreasing function of a "time" operator T that is canonically conjugate to L (MISRA, 1978),

$$M = M(T), \quad i(LT - TL) = I. \quad (8)$$

As M is a positive operator it can be written as a product of an operator A and its adjoint A^+ ($M = A^+ A$). Then, when the inequality in (7) holds, the quantity $\tilde{\varrho} = A \varrho$ obeys a manifestly irreversible equation

$$i \frac{\partial \tilde{\varrho}}{\partial t} = \Phi \tilde{\varrho}, \quad \Phi = A L A^{-1}, \quad \frac{1}{2i} (\Phi - \Phi^+) \leq 0. \quad (9)$$

The last of the relation above means that for $t > 0$, $\tilde{\varrho}$ evolves according to a (contractive) semigroup and not a unitary group as does ϱ . Such

⁵ For classical systems, more precise conditions are known (MISRA, 1978). The spectrum of L is related to the ergodic properties of the system and a necessary condition for a non-trivial M is mixing while a sufficient one is a K -flow.

a property leads to a Markov stochastic process, provided that $\tilde{\varrho}$ is a probability density (or a density matrix).

The equations (7) and (8) do not determine M , and consequently Λ , uniquely. To insure that equations (9) defined a Markov process we need to impose on Λ some supplementary conditions. One of them is that $\Lambda\varrho$ is positive and normalized to unity, so that $\tilde{\varrho}$ admits a probabilistic interpretation. Another one is that equilibrium ensembles are invariant under this transformation and thus stationary solutions of equation (9). There is not yet a proof that these requirements may be satisfied in general. Recently, however, it was possible to show that this is the case for a class of highly unstable mechanical systems, the so-called Bernouilli systems (MISRA, PRIGOGINE and COURBAGE, 1979a, 1979b).

It is remarkable that the instability of the flow that is basic in justifying the need of a probabilistic description, eventually leads to a theory of invertible but non-unitary transformations that relates reversible and irreversible evolutions. In other words, irreversibility is an inherent property of classes of dynamical systems and not the result of approximations.

Although the quadratic form for Ω suffices to settle the question of the existence of decreasing functionals of the state, other non-linear functionals may be used. In particular, for Markov processes any concave functional shares this property. Thus if we ask the entropy to be additive, the choice

$$S_m = -k \int dp dq (\Lambda\varrho) \log(\Lambda\varrho) \quad (10)$$

may be appropriate. It increases monotonically and at equilibrium reduces to the usual expression of Gibbs.

5. Concluding remarks

The common feature of dynamical systems in which irreversible phenomena take place is the instability of trajectories. In any finite region of the phase space one finds rapidly diverging trajectories or of qualitatively distinct type. Descriptions in terms of trajectories or ensembles are equivalent if we can pass from one to the other by a physically admissible procedure. We may consider a point in phase space as the limit of an ensemble but such an idealization is acceptable only if the motion is sufficiently stable. Otherwise only a statistical description is physically meaningful.

For dynamically unstable systems, a non-equilibrium entropy may be defined on the microscopic level. Once such an increasing functional is introduced, past and future have a well defined meaning. But now one must be cautious because a decreasing functional can be defined as well. There is no obvious answer to the difficult question how to choose the "proper" functional. As it is discussed elsewhere (PRIGOGINE, 1980), the distinction between past and future is a kind of a "primitive" notion that precedes physical theory.

While the existence of irreversible processes in dynamical systems is a problem that can be formulated without appealing to cosmological arguments, the possibility of performing observations cannot be dissociated from the fact that the environment is not in equilibrium. However, questions as to the meaning of irreversibility on a cosmic scale where gravitation plays an essential role are extremely difficult and no reasonable theoretical scheme has been advanced until now.

References

- ARNOLD, V. I., and A. AVEZ, 1968, *Ergodic problems of classical mechanics* (Benjamin, New York)
- BOLTZMANN, L., 1872, Wien. Ber., vol. 66, p. 275
- BOLTZMANN, L., 1896, 1898, *Vorlesungen über Gastheorie* (J. A. Barth, Leipzig). English transl. by S. G. Brush: *Lectures on gas theory* (University California Press, Berkeley, 1964)
- BRUSH, S. G., 1976, *The kind of motion we call heat*, vols I, II (North-Holland, Amsterdam)
- CLAUSIUS, R., 1865, Annalen der Physik (2), vol. 125, p. 393
- EINSTEIN, A., 1917, Phys. Zeitschrift, vol. 18, p. 121
- FORD, J., 1973, Adv. Chem. Phys., vol. 24, p. 155
- GIBBS, J. W., 1902, *Elementary principles in statistical mechanics* (Yale University Press, New Haven; reprinted by Dover Publ., New York 1960)
- GLANSDORFF, P., and I. PRIGOGINE, 1971, *Thermodynamic theory of structure, stability and fluctuations* (Wiley-Interscience, London)
- GRAD, H., 1958, *Principles of the kinetic theory of gases*, in: *Handbuch der Physik*, vol. 12, p. 205 (Springer-Verlag, Berlin)
- GRECOS, A. P., 1978, in: *Frontiers of theoretical physics*, ed. F. C. Auluck, L. S. Kothari and V. S. Nanda (Macmillan, New Delhi)
- LEBOWITZ, J. L., and O. PENROSE, 1973, Physics Today, vol. 26, p. 21
- LOSCHMIDT, J., 1876, Wien. Ber., vol. 73, p. 139
- LOSCHMIDT, J., 1877, Wien. Ber., vol. 75, p. 67
- MISRA, B., 1978, Proc. Nat. Acad. Sci. (USA), vol. 75, p. 1627
- MISRA, B., I. PRIGOGINE, and M. COURBAGE, 1979a, Proc. Nat. Acad. Sci. (USA), vol. 76, p. 3607
- MISRA, B., PRIGOGINE, and M. COURBAGE, 1979b, *Physica*, vol. 98A, p. 1

- MOSER, J., 1973, *Stable and random motions in dynamical systems* (Princeton University Press, Princeton)
- VON NEUMANN, J., 1955, *Mathematical foundations of quantum mechanics* (Princeton University Press, Princeton)
- PLANCK, M., 1900, Verhandl. Dtsch. Physik. Ges., vol. 2, p. 237
- POINCARÉ, H., 1893, Revue Métaphys. Morale, vol. 1, p. 534
- PRIGOGINE, I., 1978, Science, vol. 201, p. 777
- PRIGOGINE, I., 1980, *From being to becoming* (Freeman, San Francisco)
- PRIGOGINE, I., and A. P. GRECOS, 1979, in: Problems in the foundations of physics, ed. G. Toraldo di Francia (North-Holland, Amsterdam)
- SCHÖPF, H. G., 1978, *Von Kirchhoff bis Planck* (Akademie-Verlag, Berlin—F. Vieweg, Braunschweig)
- SINAI, Ja. G., 1967, in: Statistical mechanics foundations and applications, ed. T. Bak (Benjamin, New York)
- WHITEMAN, K. J., 1977, Rep. Prog. Phys., vol. 40, p. 1033
- WIGHTMAN, A. S., 1971, in: Statistical mechanics at the turn of the decade, ed. G. D. Cohen (Dekker, New York)
- ZERMELO, E., 1896a, Annalen der Physik, vol. 57, p. 585
- ZERMELO, E., 1896b, Annalen der Physik, vol. 59, p. 793

PROBABILITY IN CLASSICAL AND QUANTUM PHYSICS

YU. V. SACHKOV

Moscow, U.S.S.R.

1. Introduction. Spectre of opinions

The fundamental character of probability ideas and research methods in modern physics has been widely recognized. Significantly, the penetration of physics deep into matter and the discovery of the structure and properties of substance (classical statistical physics) and then of atoms and microprocesses (quantum theory) has been essentially based on probability concepts. In recent years probability methods have been intensively introduced into the theory of quantum fields and their further development is considered promising there. (See SIMON, 1974.) At the same time, there are considerable differences of opinion on the foundation of probability in physics. These differences are especially conspicuous when it comes to the comparative analysis of the foundations of probability in classical and quantum physics. Thus, for example, FEINMAN and HIBBS (1965) write that when transition is made from classical physics to quantum mechanics the very concept of probability does not change. In both cases probability is described through the notions about the relative frequencies of events being analysed. Only the methods of calculating probabilities change radically. FOCK (1967, p. 173) and OMELYANOVSKY (1973, p. 120) maintain, on the contrary, that probabilities in quantum mechanics differ radically, in their very meaning, from those in classical theories.

C. VON WEIZSÄCKER (1973, p. 334) holds a radical view that "quantum theory is nothing else but the general theory of probabilities".

A widespread point of view is expressed in the following statement of JAUCH (1975, p. 2): "The probabilities which occur in classical physics are

interpreted as being due to an incomplete specification of the systems under consideration, caused by the limitations of our knowledge of the detailed structure and development of these systems. Thus these probabilities should be interpreted as being of a *subjective* nature."

"In quantum mechanics this interpretation of the probability statements has failed to yield any useful insight, because it has not been possible to define an infrastructure whose knowledge would yield an explanation for the occurrence of probabilities on the observational level. Although such theories with 'hidden variables' have been envisaged by many physicists, no useful results have come from such attempts. We consequently take here the opposite point which holds that the probabilities in quantum mechanics are of a fundamental nature deeply rooted in the objective structure of the real world. We may therefore call them *objective* probabilities."

The spectre of opinions on the significance of probability in classical and quantum physics can be further enlarged. What does the variety of opinion mean? Can we lend it internal unity?

2. Classical physics: probability and chance

Ideas of classical statistical physics were elaborated on the basis of atomistic notions about the structure of substance. Classical physics was most successful in the analysis of the structure and properties of gases. Liquid and solid bodies started to be analysed by statistical methods much later and in the "involved" form. Gases are typical and most characteristic objects of investigation by classical statistical physics, and we take this into account in our general conception of probability.

Historically, the development of classical statistical physics can be represented as the extension of the field of application of ordinary mechanics, as a simple transition to the analysis of mechanical systems consisting, in practical terms, of an unlimited number of particles.

These problems could not be solved by direct methods of mechanics, as they appeared insuperably complex in this formulation. Methods had to be changed. Already Maxwell was well aware that as the molecular kinetic gas theory was being worked out a transition was made from strictly dynamical methods of mechanics to the methods of probability theory. How was this transition to be understood and appreciated?

Initially, during the emergence of classical statistical physics this transition was regarded as a forced flank manoeuvre, as a clear simplification

of research methods. Correspondingly, assertions were made that probability in classical physics is consequent upon our incomplete knowledge of the appropriate systems, upon their incomplete description. But far from all researchers were satisfied with such assessment of probability. Many held that the introduction of probability theory's ideas and methods into physics was of fundamental importance. Correspondingly, new foundations of probability in physics were searched for. The analysis of the role and importance of probability in classical physics is directly dependent on how the question is answered about the main task of classical statistical physics. In the course of research the very ideas about the main task of statistical theories were improved. At present it is a widespread view that the basic task of statistical mechanics has always been "the elucidation of the *relation* between the microscopic, molecular description and the macroscopic description of the physical phenomena" (UHLENBECK, 1973, p. 501).

Development of the statistical gas theory was preceded on the one hand by the creation of the foundations of gas thermodynamics, i.e. macroscopic (independent of atomistic notions) theory of gases, and on the other hand, by the elaboration of the theory of movement of simple objects, i.e. of classical mechanics. Statistical theory made it possible to come to an original synthesis of two basic and independent trends of investigation of corresponding physical systems, i.e. the trend moving from the properties of a system as a whole to the properties of its elements and the trend moving from the properties of elements to the integral properties of systems. The probability theory was that mathematical apparatus which allowed for such synthesis. Correspondingly, the significance of probability in classical physics is determined above all by the fact that it is a structural characteristic of a definite class of systems, gas-type systems. Under structure we understand above all the character of interconnection, the laws of interaction and mutual determination of elements within a certain whole. Knowledge of the structure of a particular system formation ensures us a transition from the knowledge of integral properties of systems to the properties of elements and back. The connection between probability and the category of structure finds its expression in the fact that the basic and central concept of all applications of probability theory in classical physics is that of probability distribution. Distributions express the existence of inner orderliness of corresponding systems and make it possible to describe both the elements and integral properties of these systems.

To characterize the structure of systems described by probability methods it is of crucial importance to analyse this structure's specifics. It is common knowledge that there are systems and systems, and this fact is manifested in their different structures. The systems' specifics can be brought out through an analysis of the specific relationships of their elements. In gas-type physical systems these specifics are expressed through such notions as independence, irregularity, indeterminacy, "autonomy", and so on. In the philosophical language, these systems' specifics can be expressed in a generalized way through the category of chance.

That is why we do not agree with assertions that probability in classical physics results from our incomplete knowledge of systems in question. Such assertions imply that probability theory's ideas and methods in physics have no independent value but only appear as an approximation to mechanics' rigorous analytical methods. But this is at variance with the actual state of things. The main job of statistical theories in classical physics, as we have seen, is essentially different: it is to disclose interconnections between levels characterizing the internal structure of the physical systems under study. Methods of ordinary mechanics were clearly incapable of coping with this job. Introduction of probability in physics has essentially enriched and broadened the whole mode of physical thought, which had earlier been based on mechanical notions and conceptions.

3. Quantum theory: its potentialities

The culmination point in applying probability ideas in physics was the development of quantum mechanics. As distinct from classical physics the fundamental character of probability ideas in quantum physics was in fact recognized on a fairly wide scale. This is explained by changes in setting the main task of research: in quantum theory probability methods are used above all for cognizing the properties and laws (regularities) of individual, separate microparticles. Transition from the study of systems formed from assemblies of particles to the study of separate particles speaks of the exceptional flexibility and fruitfulness of probability methods. This has been made possible through essential changes in the methods of postulating probability ideas: in quantum theory microparticles' states are expressed through special characteristics, above all wave functions.

There is an extensive literature and widely varying opinion on the foundation of probability in quantum physics. In our view, analysis of these questions hinges on the disclosure of the logical structure of quantum

theory. It is of much importance for the analysis of quantum mechanical knowledge that its concepts are divided into levels or classes. The first class is composed of "directly observables", as it were, i.e. positions and momentum. In theory they are regarded as typically accidental (in theoretical probability sense) magnitudes. The second class is comprised of quantum numbers of the spine type. Differences between these classes of concept consist above all in the "degree of proximity" to what is directly given by physical experiment. The former classes express primarily the outward, superficial characteristics of microobjects, while the latter deeper, internal ones. The former enable us to individualize quantum processes, the latter have a generalized character. The former gravitate by their nature to classical concepts, the latter above all express the specifics of quantum phenomena. The former continuously change the latter are more stable. The former are more connected with appearance, the latter with essence. These differences between classes of concepts can be described as those of a logical nature. Of course, quantum processes can be fully expressed in theory when concepts of both classes are used.

It proved possible to synthesize the magnitudes relating to different logical levels within a single theory on the basis of probability notions. Description of microparticles' states attached prime importance to the second class concepts, i.e. quantum numbers. These parameters are determined sufficiently unambiguously and underlie a quite rigorous characterization of each type of elementary particles. At the same time postulating of these parameters does not unambiguously determine the values of first class parameters, but determines the whole field of the latter's possible manifestations. A further and direct analysis of experimental situations is necessary in order to define the values of the first class units.

Quantum mechanics is sometimes said to be a science of the potentially possible in the world of microprocesses. But to say this is not to express the whole truth. In considering the spectrum of microobjects' possible behaviour quantum mechanics makes it possible to state the existence of certain orderlinesses and regularities in the "mass" of these possibilities, and its major statements are in fact based on the existence of such orderlinesses. It is very essential, moreover, that the regularities themselves in the spectrum of possibilities are conditioned by the "in-depth" properties of microobjects. The latter are defined in theory with sufficient unambiguity and expressed through the magnitudes of the second (logical) class.

What has been said above prompts the conclusion that the significance

of probability in quantum physics lies above all in the fact that it allows for a study and theoretical expression of the properties and laws of objects having a complex, "two-level" structure. The basic meaning of probability lies precisely in this connection with structure and in methods of expressing it. The peculiarities of this structure are such that the dependence between concepts belonging to different levels is already expressed not only on a coordination plane, but also through subordination. At the initial level the dependences between the values of characteristics also incorporate certain features of indeterminacy, independence, and "autonomy". At the "in-depth" level, however, the connections and dependences between characteristics are fairly unambiguous.

4. Conclusion. Development of views on probability

Let us consider again the statements on the role and meaning of probability in physics that we cited at the beginning of this report. When Feynman and Hibbs maintain that the concept of probability did not change when the transition was made from classical to quantum physics they interpret probability at the initial empirical level. A frequency interpretation of probability operates at the level of "direct observations", and in this sense transition to quantum mechanics has not essentially changed the interpretation of probability. But the meaning of theoretical ideas does not only consist in the description and coordination of direct experimental data. The general scientific concepts and categories that form the "core" of theoretical concepts express a definite cross-section of material reality. The main difficulties, and discussion on the nature of probability, mainly arise at the level of theory, where probability becomes an essential element of the abstract, generalized model of the material world, the model that evolves together with the development of knowledge. In this context the statements of Fock and Omelyanovsky, and also of Jauch, are interesting, because they emphasise essential changes in the interpretation of probability at the theoretical level when transition has been made from classical to quantum physics. During this transition probability has altered and enriched the general picture of the world, and correspondingly changed itself. C. von Weizsäcker's statement draws attention to the significance of feedback in the relations between probability and its applications. If it is recognized that probability is incorporated precisely into quantum physics in a most natural (immanent) way it follows that probability can be deeply understood by analysing the

structure of quantum mechanical knowledge. The further development of interpretation of probability in physics lies through the understanding and generalization of the situation in quantum theory.

The development and application of theoretical probability ideas and methods is on the main line of the development of modern science. They have lent essential flexibility to theoretical thinking. They also served as the basis for a certain synthesis of the continuous and discrete, stability and changeability, rigorous determinacy and independence, elementariness and wholeness, i.e. helped to disclose and reveal a deep inner dialectic of the processes of the material world. At the same time the ongoing development of science increasingly brings out the limitations of probability-based statistical methods. They make themselves felt, for example, in the solid body studies, but are especially acute in the studies of living systems.

References

- FEINMAN, R. P., and A. R. HIBBS, 1965, *Quantum mechanics and path integrals* (New York McGraw-Hill Book Company, New York)
- FOCK, V. A., 1967, *Quantum mechanics and the structure of matter*, in: *Struktura i formy materii* (Structure and forms of the matter, Russian) (Moscow)
- JAUCH, J. M., 1975, *The quantum probability calculus*, Fundamenta scientiae, Seminaire sur les fondements des sciences, No. 27 (Université Louis Pasteur, Strasbourg)
- OMELYANOVSKY, M. E., 1973, *Dialektika v sovremennoi fizike* (Dialectics in modern physics, Russian) (Moscow)
- SIMON, B., 1974, *The $p(\varphi_a)$ Euclidean (quantum) field theory* (Princeton University Press, Princeton)
- UHLENBECK, C. F., 1973, *Problems of statistical physics*, in: *The physicists' conception of nature* (Dordrecht, Holland)
- VON WEIZSÄCKER, C., 1973, *Probability and quantum mechanics*, The British Journal for the Philosophy of Science, vol. 24, No. 4

THE SCOPE AND LIMITS OF SCIENTIFIC CHANGE *

DUDLEY SHAPERE

The University of Maryland, College Park, Maryland, U.S.A.

Most philosophies of science have supposed that there is something about the scientific enterprise that is either presupposed by that enterprise, and is thus immune to revision or rejection, or else is discovered in the course of that enterprise, and is seen, from that point on, to be immune to revision or rejection. Let us call such beliefs—ones which are allegedly immune to revision or rejection—*necessary claims* about science. They need not be “propositional”, in the sense of being overt claims about the way things are. Indeed, few philosophers any longer maintain that there are any such necessary propositional claims, the nearest survivor being the view that there are “observational” statements which, insofar as they are truly observational, cannot be modified or rejected. Rather, the focus of modern philosophies of science has been on such allegedly necessary claims as the following: that there is a method, “the scientific method”, by application of which knowledge about the world is obtained, but which, once discovered (by whatever means), is in principle not subject to alteration in the light of any beliefs arrived at by its means; that there are rules of reasoning—rules, for example, of deductive or inductive logic—which are applied in scientific reasoning, but which can never be changed because of any scientific results; or that there are “metascientific”

* Research for this paper was completed during a sabbatical year at the Institute for Advanced Study, Princeton, New Jersey, in 1978–79. The paper is an extension of the more general view developed in *The character of scientific change*, also written during that year, and to appear in: *Scientific Discovery, Logic and Rationality*, ed. T. Nickles (Reidel, Dordrecht). I am indebted to a great many people for discussions leading to the ideas expressed in these papers: a long list of acknowledgments is given in the paper just mentioned. For discussion of ideas specific to the present paper, I thank particularly John Bahcall, Lindley Darden, Richard Rorty, Neal Snyderman, and Morton White. I also thank the National Science Foundation for support (under Grant SOC 76-19496) of earlier phases of the research leading to this paper.

concepts, like ‘evidence,’ ‘theory,’ ‘explanation,’ which are used in talking about scientific concepts, claims, and arguments, which have meanings which are wholly independent of the specific content of ongoing science, and which, collectively, define what science is and always will be.

Views of these sorts, maintaining that there are necessary components or presuppositions of the scientific enterprise, have faced enormous difficulties in the past. Apart from specific objections to specific variations on the theme, the general view that there are necessary claims in or about science seems to go against the trend of science in the past century or so, where we have witnessed the successive overthrow of one alleged “necessary” claim after another. More recently, historians of science and historically-minded philosophers have argued that deep change is characteristic of the knowledge-seeking enterprise throughout its development. They argue that such change extends not merely to profound alteration of our substantive beliefs about the world, but also to conceptions of the goals of science, the demarcation between science and non-science, the distinction between legitimate and illegitimate scientific problems, the methods of science, the standards of adequacy and acceptability of scientific solutions, and, in general, to everything that is a constituent of the scientific reasoning-process. Indeed, one might summarize the view proposed by these historians and philosophers as asserting that there is nothing in scientific reasoning, not even conceptions of “reasoning” itself, that is in principle immune to revision.

There is much that is appealing in this view: besides being more in accord with the history of science, particularly its more recent history, it promises to liberate us from the last shackles of apriorism and essentialism. It promises to provide us with a view of science that is uncompromisingly empiricist, in that it suggests that not only do we learn about the world through experience, but we also learn how to learn and think about it in same way.

And yet despite these appealing aspects, the view that there are no necessary claims in science faces severe difficulties. Among the many such difficulties is the following. Suppose we try to maintain that there is absolutely nothing sacred and inviolable in science—that *everything* about it is in principle subject to alteration. Then included among the things that can change are standards or criteria of what it is to be a “good reason” for change. But then how could criteria of rationality themselves be said to evolve rationally, unless there are higher-level standards or criteria of rationality, themselves immune to revision, in terms of which

changes of lower-level criteria of rationality could be judged to be rational? There thus seem to be only two alternatives: relativism, in which there is no real ground (no ground, that is, other than the decree by fiat of a triumphant community) for saying that there is "progress" in science, that one body of scientific beliefs is better than another; or else a presuppositionist theory according to which there is something, of the sort that can serve as a standard or set of standards or criteria for scientific rationality and progress, which is immune to the vicissitudes below, and which serves as the ultimate arbiter of those lower-level scientific disputes.

It is readily apparent that this difficulty is a fundamental one, and perhaps accounts in part for the reluctance of so many philosophers in the past to try to develop a theory according to which there are no necessary scientific claims. In this paper, however, I will argue that the difficulty can be overcome through a proper understanding of what is involved in something's being a "reason" in science. It will, of course, not be possible to do full justice to this important question in the brief time available today. Nevertheless, I hope the general lines of my argument will be clear.

My point of departure is the intuition that, in any argument concerning a subject-matter, those considerations will be relevant as reasons that have to do with that specific subject-matter. The question, "What does that have to do with the subject we're discussing?" is a challenge that the consideration adduced by our opponent is irrelevant, that it does not constitute a reason either way in our dispute. If this intuition sounds somewhat tautologous, let me at present merely caution that it will help us bring out important aspects of what happens in scientific reasoning. I will in any case return to discussion of its basis later. For the present, let me turn successively to two points: first, what is involved in being a scientific subject-matter, and, following that, what is involved in "having to do with" a subject-matter in science.

The science of a particular epoch can be seen as the investigation of various areas or domains. A "domain" can be defined roughly, for present purposes, as a body of information which is problematic in certain respects, and the items of which we have reason to believe are related in the sense that a unified account of them (with respect to their problematic aspects) can be expected. Domains, in this sense, can be as broad as the subject-matters of fields like electromagnetism, genetics, or organic chemistry, or as narrow as the specialized interests of individual research workers. I have discussed the mechanisms of domains in some detail in

a previous paper¹. For present purposes, what is important about them is that the division of science into domains or subject-matters is not something given by experience in any immediate way, but is something acquired through painstaking investigation, and subject to further alteration. The relationships involved are discoveries, the fruits of accumulated knowledge. That there is a subject of electricity to study had to be found out; that it constitutes a unified subject-matter and not several distinct ones (static electricity, current electricity, animal electricity, etc.) had to be established, by careful experiment and argument, by Michael Faraday. That that domain could be unified with that of magnetism was learned through a long and tortured process of investigation. Domains can be split as well as unified. But overall, science attempts to make clear and precise the various interrelationships between the items it studies, and it is indeed a mark of highly sophisticated science that those interrelationships are highly articulated.

Nor is the association of "items" into domains merely a matter of grouping together independently-describable information. On the contrary, reformulation of the language in which we talk about items often accompanies their grouping or regrouping into subject-matters or domains. Chemistry in its modern sense became a subject when older vocabulary for talking about matter—vocabulary based primarily on the sensory appearances of substances—was replaced by a new descriptive language in which substances were named in accordance with the elemental substances of which they were compounded. The possibility of that reform of nomenclature underscores the role of prior knowledge (or belief) in the formation of domains and their description. In the case of the eighteenth-century chemical reform, a number of factors entered in to make the reform feasible: the possibility of analysis into constituents (rather than those alleged constituents being created by the process of analysis); the idea that knowledge of composition gives understanding of matter; the concept of an element as a breakdown product, and the discovery that

¹ SHAPERE, D., 1974, *Scientific theories and their domains*, in: The Structure of Scientific Theories, ed. F. Suppe, pp. 518–565 (University Illinois Press, Urbana). Many of the specific cases discussed in the present paper have been developed more fully in earlier papers; for example, the case of the chemical revolution is dealt with (though in a somewhat different way) in: *The influence of knowledge on the description of facts*, in: F. Suppe and P. Asquith (eds.), PSA 1976, East Lansing, Philosophy of Science Association, 1977, pp. 281–298. Other cases are discussed at length in as yet unpublished but forthcoming papers.

there are a relatively small number of breakdown products common to a wide range of chemical processes; the concept of weight as being centrally relevant chemically; and Lavoisier's oxygen theory of acids, metals, and calxes. With the further development of an area of science, the formulation of domains tends to become more and more built on previous knowledge or well-grounded belief.

The relations between the descriptive language of domain items and the "observation-language" in terms of which hypotheses about the domain items are tested is not straightforward and simple. Hence it is necessary here to emphasize that that observation-language, too, tends, in sophisticated science, to rest on a vast store of prior well-founded beliefs. As an example, consider the claim, universally made by astrophysicists since the early 1960's, that it is now possible to make direct observations of the center of the sun. What is involved, of course, is a knowledge of the behavior of the neutrino—specifically, but only in part, of processes in which neutrinos are emitted, of the fact that neutrinos are weakly-interacting and therefore can be expected to pass freely to us, without interruption or interference, from their origin through the body of the sun and interplanetary space, and of the kinds of receptors appropriate for intercepting such neutrinos. And it is prior physical knowledge that specifies what counts as an "appropriate receptor", the ways in which information of various types is transmitted and received, the character of interference and the circumstances and even the statistical frequency with which it occurs, and even the types of information there (fundamentally) are. The physics of the present epoch, for example, makes such specifications for a wide range of circumstances through the whole body of well-founded belief about elementary particles, their decays and interactions, and the conservation principles which govern such processes. In particular, for example, knowledge of the cross-sections for such particles, the probabilities of their interactions with other particles (or decays of individual particles) in given environments, contributes to specifying the notion of "interference" or "interruption". In particular problems far more enters into specifying such notions as that of a "receptor" than just what is contained in elementary particle theory ("observation" is not laden merely with "theory"); for example, in the solar neutrino case, knowledge of the environmental conditions (pressure, temperature, opacity) in the interior of the sun also plays a role, as does knowledge about various instruments.

So the formation of domains or subject-matters, and their description,

and the language in which we express observational tests of hypotheses about domains, are highly dependent on “background information”, on accumulated knowledge or presumed knowledge which is brought to bear on that formation and description. Analogous considerations and cases could be given to show how other aspects of the development of science are also shaped by background belief: the problem-structure of the science of a given epoch, for example, or the range of possibilities envisioned at that epoch. But what about that “background information” itself? What constraints govern its conception and employment? What justifies the claim that certain beliefs “have to do with” a subject-matter and can enter into considerations about it?

In relatively early and unsophisticated stages of science, the line between “scientific” and “unscientific”, between the scientifically relevant and irrelevant, is not clear, or is even nonexistent: there is, at that time, no basis for exclusion of certain sorts of considerations as scientifically irrelevant. (There may, indeed, be grounds for supposing them to be relevant, even though later science will exclude them.) In the seventeenth and early eighteenth centuries, there was no clear line between what we today would distinguish as “scientific” and “theological” considerations. Newton, indeed, believed that there were at least three ways in which the laws of physics implied the necessity of God’s interference in the world: first, in order to keep the motions of bodies from dying down in the face of momentum loss through impacts; second, in order to preserve the order of the solar system against the disruptive effects of gravitational perturbations; and third, in order to resolve the inconsistency of a universe in which, if matter were distributed finitely through space, all bodies would fall to their mutual center of gravity, but in which, if matter were distributed infinitely through space, all gravitational interactions would be cancelled. Yet by the middle of the eighteenth century, confusions about *vis viva* and momentum had been resolved, and Newton’s first reason had dissipated; and Laplace showed, or at least claimed to show, that gravitational forces between planets are self-correcting in the long run. Apocryphal or not, the story about Laplace’s reply to Napoleon’s query about the place of God in Laplacean science (“We have no need of that hypothesis”) captures the fact that, to Laplace and the increasing majority of his physicist successors, theology had been excluded as irrelevant to science. (The third Newtonian point, the Bentley–Seeliger paradox, was neglected until the twentieth century.) In a similar vein,

Keplerian questions about astrological influences, or about why there are exactly six planets—questions Kepler himself admitted because they were equally as “geometrical” (in his interpretation of that concept) as questions about the shapes of planetary orbits and about the relations between orbital speed and distance of a planet from the sun—came to be excluded by the success of the view that matter exerts causal influence only by impact and (possibly) attraction and repulsion at a distance. These examples show how certain types of considerations come to be excluded as scientifically irrelevant; but there are also, sometimes, cases involving the introduction of radically new types of considerations. Thus the work of Gauss and Riemann made it possible to consider physical space as having intrinsic characteristics without supposing it to be embedded in a higher-dimensional space, and further, to allow those characteristics to vary from point to point and from time to time. Newton and Leibniz had rejected the Cartesian idea that matter can be understood in purely geometrical terms precisely on the ground (among other reasons) that it was inconceivable for space to have such characteristics.

In general, then, as science develops, successful beliefs produce constraints on what can count as scientifically relevant, sometimes broadening, sometimes narrowing, the range of the scientifically relevant. The line between the scientific and the non-scientific—between what can count as a scientific consideration and what cannot—is thus an acquired characteristic of scientific inquiry; it is not innate, essential to and definitory of the scientific enterprise itself. In its earlier phases, science relies on considerations that may later be excluded, or fails to rely on considerations that will later be alleged to be relevant. Metaphysical, theological, political considerations all have played roles in the development of scientific ideas, as have otherwise ungrounded analogies or symmetry considerations. Such considerations have later been found to be irrelevant, in the light of what we have learned about the world; or, as in the case of analogies, they have been transmuted into or foresworn in favor of relations which are supported by a background of successful beliefs. Thus stellar classification was originally (last quarter of the nineteenth century) based purely on the colors of stars, and theories of stellar evolution on the analogy of stellar colors with the progression of colors in heated objects. The success of such an approach (in a highly modified form) has led to the conviction that there is *evidence*, not mere analogy, in favor of the view that astronomical objects of all sorts are thermodynamic entities. And

symmetry considerations have passed from being purely aesthetic demands to being very precise, integral, and testable constituents of elementary particle theory.

This is not to say that science, or any particular area thereof, has wholly dispensed with the need for now-recognizably non-scientific or questionably justified considerations such as analogies. Even in so paradigmatically advanced an area as particle theory, Yukawa relied on analogy with the extant quantum theory of electromagnetic interactions in constructing his exchange-particle theory of strong interactions. And later physicists appealed to non-Abelian gauge theories of the Yang-Mills type, successful already in quantum electrodynamics, as a model for construction of theories of the weak and later of the strong interaction. (It is true that, in these cases, certain difficulties had to be overcome before the analogy could prove applicable.) And many physicists continue to hope for a similar theory of gravitation.² Yet the clear hope of science is to exclude the need for such considerations: to develop its body of well-founded beliefs to such an extent that it will include within itself the grounds for all such reasoning. Science attempts, so to speak, to *internalize* the considerations that are to count in its further deliberations. That is a procedure that has proved successful in the past, so successful that it has become a normative principle governing what is to count as a "reason" in science.

But this brings us back to our starting-point. For what we have found is that science attempts, as far as possible, to develop in such a way as to exemplify the intuition that what is relevant as a reason in science must have to do with the subject-matter at hand. The explicit development of domains is a direct instance of this intuition: the attempt to formulate, as clearly and unambiguously as possible, what the subject-matter of an area of investigation is. Similarly, the evolution of the distinction between scientifically relevant and irrelevant considerations, a distinction made in the light of our successful beliefs, works in the direction of making explicit and unambiguous which beliefs may be brought to bear in the further building of our scientific conceptions. Among those successful beliefs, some will be established, to some degree at least, as relevant

* It is also true, however, that the gauge-theoretic approach *might* ultimately be discarded in favor of some other which might (for example) be more conducive to a unification of weak, electromagnetic, and strong interactions with gravitation. It has been suggested by Misner, for example, that formulations in terms of harmonic mappings should be explored in this connection. (C. MISNER, 1978, *Physical Review D*, vol. 18, pp. 4510-4524).

to specific domains, or to specific instrumentation which may in turn provide the bases for tests of specific hypotheses about specific domains. Thus, what I have spoken of as the "internalization of considerations" in science is simply a specification of what is to count as a "reason". And again, it is fully in accord with the intuition about reasons with which I began—namely, the intuition that those considerations are to be considered relevant as reasons that have to do with the subject-matter at hand.

What is the status of that intuition itself? It, too, has been acquired in the course of inquiry; it is not the product of a purely intellectual analysis of the essence of "reasoning". The recognition of the need to delineate our subject-matter clearly and sharply, and to specify as exactly as possible what is and is not relevant to that subject-matter, has not been present in the knowledge-seeking enterprise from its inception. Science has gradually become aware of it, and seen its success as an approach. And that success in turn has elevated it to the status of a guiding principle of scientific inquiry.³

In short, then, science builds on the basis of its successful beliefs. But "success", while in general necessary, is not (in general) sufficient to qualify an idea for being "built into" further scientific development—into, for example, the formulation of new domains, or the new formulation of old domains, or the introduction of new possibilities. For a belief can be successful while there still exist specific reasons for supposing that it cannot be true, that things could not really be that way. Such beliefs we call "idealizations", and I have analyzed their role in the scientific enterprise elsewhere.⁴ In general, idealizations do not serve as "background

³ One might wish to argue that, even though people were not aware of adhering to the principle that considerations must be relevant to the subject-matter at hand in order to count as reasons, they nevertheless always adhered to it, and must have done so, because the principle is an "analytic truth". Such a claim is not, however, cogent: notions of relevance, subject-matter, and argument were far too vague in earlier times to support it without distorting the thought of those times. Nor should one claim that, even though the principle may have been discovered, it is nevertheless a necessary truth, not subject henceforth to revision or rejection. For what are counted as (or the "criteria" for identifying) a "subject-matter" and a "relevant consideration" are clearly subject to revision in the light of further scientific discovery. And I have argued elsewhere (*The character of scientific change*, Part III) that the "meaning" of the "principle" in question is exhausted by such criteria (or, for certain purposes in certain contexts, by the family of such criteria that have been developed so far in history), and that even the "principle" itself might be rejected under certain scientifically-determined circumstances.

⁴ SHAPERE, D., 1969, *Scientific theories and their domains*, loc. cit., Part IV; *Notes toward a post-positivistic interpretation of science*. Part II, in: *The Legacy of Logical Positivism*, eds.

information" for the construction of new beliefs (or methods, etc.). For that purpose (in general) there must—ideally—be no specific reason for doubting the belief. (I qualify these remarks by saying "in general", because, in certain circumstances—such as the lack of anything better to go on, or the questionability of its idealizational status—an idealization *can* be so used. I should also note that, as in my other writings, I contrast "specific reasons for doubt" with "universal or philosophical 'reasons' for doubt", the latter—like "A demon may be deceiving me", or "I may be dreaming"—applying indiscriminately to any claim whatever. Such philosophical doubts play no role in the scientific or knowledge-seeking enterprise.)

Thus a "reason" in science consists of a belief (a) which has proved successful, (b) concerning which there is no specific reason for doubt, and (c) which has been shown to be relevant to the specific domain in which it is being applied as a "reason". These characteristics hold as ideals, of course: in practice, we must rely on beliefs that have not proved unambiguously successful or unambiguously free from doubt.

As an example, consider again the investigation of stellar energy production and stellar evolution via solar neutrinos. Since the primary theory of energy production for stars in the sun's mass range is the proton-proton reaction, the highly successful apparatus of elementary particle physics is brought to bear. The applicability of particle physics, together with more specific knowledge, leads to the conclusion that cleaning fluid (C_2Cl_4) is an appropriate medium for the reception of the particular solar neutrinos being sought. By inverse beta-decay, the isotope chlorine-37 is converted to argon-37, and thus, in turn, facets of the chemistry of argon become relevant in deliberations about the solar neutrino problem.

I have thus far left the notion of "success" unexamined, and in spite of its centrality, will have to leave it largely unexamined here. I have argued extensively elsewhere that we learn what success is in science. Within the very general context of "dealing with experience", many different conceptions of the goals of science, and of what constitutes "success" in "dealing with experience", are found in the history of science. The chemical revolution of the eighteenth century, for example, carried with

P. Achinstein and S. Barker (Johns Hopkins University Press, Baltimore), 1969, pp. 115-160; *Natural Science and the future of metaphysics*, in: Methodological and Historical Essays in the Natural and Social Sciences, eds. R. Cohen and M. Wartofsky (Reidel, Dordrecht, 1974), pp. 161-171. These ideas are incorporated and extended in *The role of conceptual devices in Science*, in process.

it a change in conception of the goal of matter-study, from (as one view among many) the idea of bringing matter to perfection to understanding matter in terms of its constituents. That change of goal brought with it changes in conceptions of what it is for a view of matter to be "successful". Standards of success are among our beliefs, and there are a variety of ways in which they can change without the assumption of a transcendent, unchanging criterion of success. For example, at some stage one standard of success may be dominant, but we find we cannot fulfil it, while another "lower" standard gets satisfied frequently and well, and people begin paying more attention to it. This was, essentially, the case in the chemical revolution.

In the light of the account I have presented, we can see why science need not appeal to a transcendent and irrevocable principle of rationality in order to account for the occurrence of rationality and progress within scientific change. For what better standards or criteria could we employ—at least when we are able—than those beliefs (and methods and so forth) that have proved successful and have not been confronted with specific doubt? In the attempt to find some basis for considering certain things to be observable, or for distinguishing between those hypotheses to consider and those not to consider, and so forth, what else should one expect to use and build on, wherever possible, if not such beliefs? No further sorts of reasons are available to us, and none further are required, in order to account for the rationality and progress of the scientific enterprise. That our reasons, or the beliefs and methods which constitute their bases, may be wrong, and are sometimes shown to be, does not invalidate the practice of using the best information (and *hence* the best reasons) we have available—"best" in the sense of having shown themselves successful, and not having become subject to specific doubt.

And thus the difficulty mentioned earlier regarding the view that there are no necessary claims in or about science, all being in principle revisable in the face of experience is, if I am right, dissipated. That objection, recall, was that, unless we appeal to transcendent standards of rationality which are immune to revision, we cannot have any way of concluding that scientific change is rational. My response has been that the reasons we have available (in a clear sense of 'reason') are used in making such judgments, and that no other considerations need be appealed to. Other difficulties remain in defending the view that scientific change is in principle pervasive and void of necessary claims; but those further difficulties must be examined on another occasion.

ASPECTS OF THEORY CONSTRUCTION IN BIOLOGY

LINDLEY DARDEN

*Committee on the History and Philosophy of Science
University of Maryland, College Park, Maryland, U.S.A.*

Philosophers, impressed by the certainty provided by deductive logic, have, in the past, demanded that science provide the same kind of certainty. But arguments to the contrary are persuasive. Duhem claimed: "Unlike the reduction to absurdity employed by geometers, experimental contradiction does not have the power to transform a physical hypothesis into a indisputable truth." (DUHEM, 1914, p. 190). Recognizing these limitations, Stephen Toulmin said: rationality need not be equated with logicality: "A man [or woman] demonstrates his rationality, not by a commitment to fixed ideas, stereotyped procedures, or immutable concepts, but by the manner in which and the occasions on which, he changes those ideas, procedures and concepts." (TOULMIN, 1972, p. x). Other recent philosophers have been concerned to understand the rational factors involved in scientific change, most notably Imre LAKATOS (1970) and Dudley SHAPERE (1974).

This paper is within this recent tradition in philosophy of science of understanding the rational means by which science changes. More specifically it is concerned with the question—how are theories constructed and how are they modified in the light of new evidence? Duhem might have been right when he claimed: "No hypothesis which is a component of a scientific theory T can ever be sufficiently isolated from some set of auxiliary assumptions or other so as to be separately falsifiable by observations." (QUINN, 1974, p. 36). But the truth of such a claim, based on the rigorous demands of certainty, does not negate the possibility of finding rational means, better or worse means, of modifying a theory in the light of conflicting evidence. This paper will argue that procedures

exist for determining which postulates of a theory¹, as opposed to others, are more likely to be in need of modification when evidence necessitates a change. In at least some cases, theory construction is a modular process with specific postulates constructed to account for specific data. Consequently, changes in that data direct the localization of the needed change to the corresponding postulates. Furthermore, as philosophers have often noted (e.g. DUHEM, 1914, p. 185), if a set of postulates is used to make a prediction about new phenomena, not previously investigated, and that prediction is not confirmed, then that set of postulates is in need of modification. But if that set is small and is independent of other postulates of the theory, then the locus and range of modifications may reasonably be judged with some accuracy.

In order to substantiate these claims for patterns of reasoning in theory construction and modification, this paper will analyze a case from twentieth century biology, the construction of the theory of the gene, from the rediscovery of Mendel's work in 1900 to the statement of the theory by T. H. MORGAN in 1926. We will see which postulates accounted for which data and then trace how a subset of the postulates were challenged and modified in the light of new evidence.

At the beginning of the field of genetics around 1900, the term "gene" had not come into usage. Instead, the theory was expressed either in terms of differences among germ cells or some kind of "factors" within germ cells. Details about these factors were added as the theory of the gene was modified and augmented. The original postulates were constructed to solve problems posed by empirical regularities, the famous ratios discovered by MENDEL in 1866 and rediscovered in 1900.

Table 1 presents Mendel's data, the domain to be explained. Items 1 and 2 present the puzzle, the major problem, that called for a theoretical

¹ A note on terminology. I am using the word "theory" for a set of "postulates." "Hypothesis" is a more flexible word which refers to a claim not yet confirmed, which could be an alternative postulate. I refer to the changing set of postulates as composing the *same* theory, which is undergoing modification as new hypotheses are proposed and tested. I have not faced the problem of whether there is one (or more) "essential" postulate(s), such that if it (they) were changed the theory would cease to be the same one and would become a different theory. I suspect "purity of the gametes" was such a postulate in the Mendelian case and the scientists tended to talk as if it were, but I have not examined that carefully. My usage departs from the way philosophers normally talk (e.g. LAKATOS, 1970) in which any change in postulates changes T_n to T_{n+1} . I prefer discussing the development of a single theory rather than discussing the proposal of a series of new theories because I think it fits the scientific usage more accurately.

explanation: how can something be present (e.g., green color in a parent), disappear for a generation (e.g., all F_1 hybrids yellow), and then reappear in a pure form (e.g., pure green in F_2)? This problem was solved by introducing a new idea, that is, postulating a theoretical entity, a hidden factor, to account for the puzzle. The factors, it was claimed, are transmitted unchanged and thus again produce pure characters. At the outset in theory construction, one needs a simple connection between the theoretical entity and empirically determinable items. One needs a way of easily inferring from empirical items to numbers, types, behaviors of the theoretical entities which are otherwise inaccessible. Theory construction in genetics satisfied this need: one factor was claimed to be associated with one independently variable trait of a character. Information about the behavior of observable characters was used to infer behavior of factors.

Table 1

| |
|--|
| <p>Genetic phenomena to be explained:</p> <p><i>Item 1.</i> In artificial breeding experiments involving crosses between varieties of animals or plants differing in the traits of one character, one trait dominates over the other in the hybrid (F_1) generation. (For example, yellow and green color are differing traits for the character of pea color; yellow \times green peas yield hybrids all of which are yellow.)</p> <p><i>Item 2.</i> When hybrids from crosses such as those described in Item 1 are allowed to self-fertilize or are crossed with each other, on an average one obtains a ratio of 3 dominants to 1 recessive in the next (F_2) generation. (For example, yellow hybrid \times yellow hybrid yields 3 yellow: 1 green.) When the recessives from the F_2 cross are self-fertilized, all offspring are recessive. (For example, green \times green yields only green.) When the dominants from the F_2 cross are self-fertilized and followed for successive generations, then one third yields pure dominants while two thirds again behave as hybrids. (For example, yellow \times yellow yields 1 pure yellow: 2 hybrid yellow.) The 3:1 ratio in the F_2 thus resolves into 1 pure dominant: 2 hybrids: 1 pure recessive.</p> <p><i>Item 3.</i> In other experiments involving crosses between varieties differing in traits of two characters, the characters behave independently, giving on an average an F_2 ratio of 9:3:3:1. (For example, yellow tall \times short green peas gives hybrids that are yellow and tall: when these are self-fertilized the F_2 gives, on an average, 9 yellow tall : 3 yellow short : 3 green tall : 1 green short.)</p> |
|--|

Once a theoretical entity has been postulated, one may then ask numerous questions about it. For example, where is it located? The genetic data about characters and their numerical relations provided no answer to this question. At this point in theory construction, one may have to turn to other fields in order to answer supplementary problems raised by the postulation of the theoretical entity. In this case, the neighboring field

of cytology provided the answer: since the germ cells (gametes) alone provide the link between generations, then the factors must be carried in the germ cells, that is, carried in the pollen and egg cells of plants and the sperm and eggs in animals.

Table 2

| | Postulates as of 1900 |
|-------------------------|--|
| Theoretical assumption | <ol style="list-style-type: none"> 1. Traits of characters are produced by factors (elements, pangens, <i>Anlagen</i>, later genes). |
| Simplifying assumption | <ol style="list-style-type: none"> 2. One independently heritable trait of a character is produced by one factor; called unit-character concept. |
| Interfield connection | <ol style="list-style-type: none"> 3. Since germ cells (i.e., pollen and eggs in plants, eggs and sperm in animals, also called gametes) are the links between generations, the factors are passed from parent to offspring in the germ cells. |
| Dominance-recessiveness | <ol style="list-style-type: none"> 4. In a hybrid formed by crossing parents that differ in two traits of a single character, there is some difference between the factors such that one dominates over the other and thus determines that the character in the hybrid resembles one but not the other of the parents. (Let A symbolize the dominant trait which appears; a the recessive which is not apparent.) |
| Segregation | <ol style="list-style-type: none"> 5. The parental factors are not modified as a result of being together the hybrid, nor are any new kinds of factors formed. 6. In the formation of the germ cells of the hybrid, the parental factors separate (segregate) so that the germ cells are of one or other of the parental types. This is called "purity of the gametes" and symbolized by saying that each germ cell carries A or a but not both. 7. The two different types of germ cells are formed in equal numbers in the hybrid. 8. When two similar hybrids are fertilized (or self-fertilization occurs), the differing types of germ cells combine randomly. $(A+a)(A+a) = AA + 2Aa + aa$; appearance 3A:1a. |
| Independent assortment | <ol style="list-style-type: none"> 9. The factors in hybrids formed from parents differing in two traits of two characters assort so that all pairwise combinations are found in the germ cells. (AB, Ab, aB, ab) 10. The four different types of germ cells are formed in equal numbers. 11. When two similar hybrids are fertilized (or self-fertilization occurs) the four different types of germ cells combine randomly. $(AB+Ab+aB+ab)^2 = \text{complicated array which in appearance reduces to } 9AB:3Ab:3aB:1ab$. |

With the theoretical assumption of hidden factors (Postulate 1 of Table 2), the simplifying assumption of one unit-one character (Postulate 2) and the interfield connection linking factors to germ cells (Postulate 3), the further postulates to give specific explanations of the domain item could be constructed.

Table 2 is my attempt to lay out more clearly than was done at the time the separable assumptions made by the theory as of 1900. Note that no explanation of Item 1, dominance-recessiveness, is given by Postulate 4: some difference, not known, between factors was claimed to be responsible. This was easily dropped when exceptions were found. Postulates 5-8 together accounted for the 3:1 ratios (Item 2) and were collectively usually called "segregation" or "Mendel's first law". However, each postulate was separable assumption that was challenged historically. Postulate 6 was often called the "essence" of segregation; the "purity of the gametes" was often cited as the essential discovery of Mendel. DARWIN in 1868, not knowing of Mendel's work, had postulated that hybrid units were formed in the hybrid. Mendel's 3:1 ratios could not easily be explained on such an assumption, but pure parental factors with no hybrid units easily explained the data. Postulates 9-11 explain the 9:3:3:1 ratios (Item 3) and were collectively referred to as "independent assortment" or "Mendel's second law" (although they were not clearly laid out as postulates separate from segregation until after 1900). Postulates 9-11 were all challenged and subsequently modified.

The theory as of 1900 was rather different from the statement of the theory of the gene given by Morgan in 1926:

The theory [of the gene] states that the characters of the individual are referable to paired elements (genes) in the germinal material that are held together in a definite number of linkage groups; it states that the members of each pair of genes separate when their germ-cells mature in accordance with Mendel's first law, and in consequence, each germ-cell comes to contain one set only: it states that the members belonging to different linkage groups assort independently in accordance with Mendel's second law: it states that an orderly interchange—crossing over—also takes place, at times between the elements in corresponding linkage groups: and it states that the frequency of crossing-over furnishes evidence of the linear order of the elements in each linkage group and of the relative position of the elements with respect to each other. (MORGAN, 1926, p. 25.)

A number of changes had been made. Dominance-recessiveness was no longer included among the postulates. Segregation had withstood all the tests to which it was subjected. Independent assortment was substantially modified by the discovery and explanation of linkage. It is too

lengthy a task for this paper to trace all the challenges and modifications in the theory. Instead we will focus only on the unit-character assumption, Postulate 2, and segregation, Postulates 5-8.

The unit-character concept, Postulate 2, was a simplifying assumption important for obtaining empirical access to the hidden factors. But the direct relation of one trait of a character produced by one factor was too simple, and working out the more complicated relations aided in both the conceptual development of the theory and extended the domain to which it applied. JOHANNSEN (1903, 1909) did experiments in which he selected traits which continuously varied in populations, e.g., the size of beans. Such "continuous" or quantitative variation was not part of the original domain of Mendelism and had been thought to need a different theory to explain it. Johannsen showed that selection was effective only in sorting out different genotypes (a term he coined) which produced the statistically varying characters of the phenotypes. With a clearer understanding of genotype and phenotype, other workers (e.g., EAST, 1910) followed up BATESON's suggestion (1902) that multiple factors interacted to produce seemingly "continuous" variation. The unit-character concept thus gave way to knowledge of more complicated relations between genes and characters and the domain of the theory was enlarged to include what had once been thought to be an exception.

The central claims of the theory are expressed by Postulates 5 and 6, namely that factors are not modified by being together in the hybrid and gametes consist of pure parental factors. These postulates underwent extensive tests with W. E. Castle as their strongest critic. Yet, after years of experiments, he finally came to accept them as a result of a crucial experiment. Since, at the outset, Castle accepted a strong connection between units and characters (Postulate 2), he reasoned in the following way: variation is found in traits which were once present together in the hybrid; thus, the units producing those traits may be inferred to vary also. By selecting individuals with one or other of the extremes of the variation, CASTLE (and PHILLIPS, 1914) claimed to be selecting modified units. H. J. MULLER (1914) provided an alternative interpretation: the multiple factor hypothesis already proposed (modification of Postulate 2) could be invoked to postulate "modifying" factors which interacted to produce the variation and which Castle was sorting out into pure lines by selection.

Castle recounted the events and issues:

Investigations with rats were made by us primarily to ascertain whether Mendelian characters are, as generally assumed, incapable of modification by selection... My own

early observations indicated that they were modifiable; and to this view I stubbornly adhered, like Morgan in his early opposition to Mendelism, until the contrary view was established by a crucial experiment. (CASTLE, 1951, p. 71.)

Experiments were carried out on rats with a black and white pattern, called "hooded" and known to segregate as a Mendelian character. In separate series of experiments, Castle selected in both directions, toward more white and toward more black.

Castle continued:

The fact of modifiability of the hooded pattern was thus firmly established, but its interpretation was still doubtful. Two interpretations were possible: (1) that the unit character, or unit factor, or gene for the hooded pattern, as it had successively been designated, was itself fluctuating, or (2) that the observed modification had been effected by a modifying influence of other genes than the gene for hooded pattern. Such hypothetical genes might be designated modifying. Their reality in other genetic material became increasingly clear from 1911 on. (CASTLE, 1951, pp. 71-72.)

With hindsight Castle was willing to give the hypothesis of modifying genes more credence than he did while opposing it. In 1919, discussing modifying genes, he said:

In some cases [the modifying genes] are known to have other functions also. Thus the gene proper of one character may function also as a modifying gene for another character. But in the majority of cases the only ground for hypothesizing the existence of modifying genes is the fact that characters are visibly modified.

As an alternative to the theory of modifying genes, the theory has been considered that genes may themselves be variable and if so, genes purely modifying in function might be dispensed with. (CASTLE, 1919, p. 127.)

Castle is here appealing to a criticism of modifying genes as being *ad hoc* addition to the theory in order to save Postulate 5, namely that genes are not modified by hybridization. He advocated, instead, dropping that postulate in favor of modifiable genes, a hypothesis that may be argued to be the simpler of the two since it fulfills the dictum not to multiply entities beyond necessity. But Muller's counter to these arguments was that the multiple factor hypothesis was to be preferred because it was consistent with previous work and did not involve the "radical" denial of the conclusion to which "all our evidence points", namely "the conclusion that the vast majority of genes are extremely constant", and change only by occasional mutation (MULLER, 1914, pp. 61-62).

Castle saw very clearly what was needed to perform a test to choose between the hypotheses:

Since the supposed "factors" of inheritance are invisible, we cannot hope to deal with them directly by experiment, but only indirectly. Our method obviously should be to eliminate all environmental factors so far as possible and also all factors of inheritance except one... But it is very difficult to apply this method to specific cases, since when variation is observed it is always possible to suppose that all factors but one have yet been eliminated. (CASTLE, 1916, p. 95.)

Despite the difficulties, Castle designed and carried out the crucial experiment. He took hooded rats which had been selected and crossed them with another race which lacked the hooded character. If the original hooded trait reappeared in the new hybrids, it was predicted, then selection had not modified the factors; instead, modifying genes were eliminated by the cross and the original factor was once again producing the hooded trait. The prediction was confirmed.

Castle accepted the results as crucial and changed his views:

The fact that the hooded gene itself had not been changed as a result of long continued selection was thus demonstrated, but, incidentally mutation had been observed to occur in the hooded gene itself in a single instance. Thus we now knew that in mammals color patterns may be modified by selection (1) through cumulation of modifying genes, and (2) much more rarely, by isolation of mutations in the particular pattern gene itself. (CASTLE, 1951, p. 72.)

Thus, Castle's tests of the central claims of the theory (Postulates 4 and 5) resulted in their retention. More evidence for the multiple factor hypothesis led to further evidence against the one unit-one character concept. Furthermore, "modifying genes" became part of the multiple factor repertoire. Referring to Table 2, Postulates 5 and 6 were tested and confirmed but those tests led to further change and development of Postulate 2, namely a change from the one unit-one character concept to the proposal of the interaction of multiple factors.

Postulates 6, 7 and 8, collectively often called *segregation*, came under fire when CUÉNOT (1905) found an exception to the 3:1 ratios that those postulates were constructed to explain. On breeding yellow mice with those having other colors, yellow was dominant. However, when the hybrids were bred, the percentage of yellow in the F_2 generation was smaller than expected by 2.55 per cent, that is, the ratios were between 2:1 and 3:1. When Cuénot bred the yellows so produced, he was unable to obtain any homozygous yellows, i.e., no pure dominants (no AA) were produced. CASTLE and LITTLE (1910) tested Cuénot's results with a larger sample and showed that the ratio was closer to 2:1. It was agreed that Postulates 6-8 were the locus of the problem, but Cuénot, Castle

and Morgan proposed different modifications in the postulates to account for this exception which occupy a scale from a radical change to little change of postulates. Morgan's entailed the most fundamental modification, a denial of the purity of the gametes in general (Postulate 6), with this case providing evidence for their impurity. CUÉNOT (1905) left Postulate 6 intact but modified Postulate 8 by proposing a selective fertilization. Castle's explanation involved the least modification: he left all postulates intact and explained away the exceptional data as an unusual case of inviability of some gametic combinations.

The differences among the proposed modifications are instructive, so we will examine these alternatives in more detail. In 1905, as a critic of Mendelism, MORGAN (1905) used Cuénnot's exceptional results to provide evidence against purity of the gametes. He proposed that the factors never segregate into pure parental forms but that the hybrid produces gametes with both dominant and recessive factors present. However, in half the gametes the dominant is latent; in the other half the recessive is latent. In future generations, Morgan predicted, the effect of the hybridization would be evident, some of the previous grandparental characters would reappear. In Cuénnot's exceptional yellow mice such appearances occurred sooner than is usually the case. Morgan, it should be noted, here is giving a theoretical explanation which accords with intuitive suspicions. The Mendelian phenomena are puzzling: is it really possible that a hybrid yellow pea can give rise to a pure breeding green strain that will never show the effects of having been produced by a yellow hybrid? Purity of the gametes entails that result. But when MORGAN (1909) tested his prediction with other strains of mice, he did not find the predicted reappearance of dominants in the recessive strains. His conclusion: "It is evident that the hypothesis failed when tested and must therefore be abandoned." (MORGAN, 1911).

Cuénnot's modification left the postulate of purity of the gametes (Postulate 6) intact and instead modified Postulate 8. Not all gametes combine randomly, he claimed; sometimes selective fertilization occurs. In this case the gametes bearing factors for yellow selectively combine with gametes bearing different factors, but not with each other. (In the symbolism used here: $(A+a)(A+a) = \text{no } AA, \text{ only } 2Aa : 1aa$.) MORGAN (1909) criticized Cuénnot's hypothesis of selective fertilization as "a conception entirely foreign to the whole Mendelian scheme. There is no evidence of selective fertilization *in this sense* known elsewhere and it seems a very questionable advantage to introduce the factor [an unfortunate

choice of word here] into the Mendelian process." (MORGAN, 1909, p. 503). Bateson and Punnett also criticized selective fertilization by counter ing that it would not even explain the exceptional ratios. Since more sperm are found than eggs, there would be sufficient numbers of "non-yellow" sperm to fertilize all the "yellow" eggs so one would expect a 3:1 ratio of yellow to other color (e.g., 3Aa : 1aa) even though no pure yellows were among the proportion of yellows. (BATESON, 1909, p. 119.)

The lack of pure yellows still required an explanation and Castle's explanation proved to be the best: the embryos formed by the mating of two gametes with yellow factors were inviable, thus one expects the 2:1 ratio that was found. CASTLE (and LITTLE, 1910) appealed to smaller numbers of young produced as evidence for this hypothesis and also pointed out that decreased viability had been found in other cases. By 1941 (e.g., MORGAN, 1914) Castle's explanation of Cuénot's results had been accepted.

Once again, the purity of the gametes and the other segregation postulates survived severe challenges to remain unmodified. Such was not the case with the independent assortment postulates (9-11) which were extensively modified as numerous cases of linkage were found. But a discussion of those modifications is too lengthy for consideration here.

A number of features of this case may be noted. The original domain resulting from Mendel's work with peas was very simple in that it consisted of results from a single species with clear cut character differences. It was a good model system for seeing the need to postulate pure units not interacting in the hybrid. Furthermore, the technique of artificial breeding provided a means of developing a line of research to test the generality of the postulates. The promise of this new theory marked the emergence of the field of genetics to carry out tests of generality. (See DARDEN, 1977, 1978 for further discussion.) But, as might have been expected, hereditary phenomena were more diverse. (As in fact Darwin had already known; Darwin did not discover Mendel's laws, it is plausible to claim, because he knew too much.) Experimental results expanded the domain, necessitating modifications in the postulates. Some were straight forward "complications", i.e., the original postulate was too simple and was expanded. The change from the unit-character concept to multiple factors is an example. In some cases, that the particular data required modification in a particular postulate was not debated. The postulate had been introduced to account for an item in the domain when that item was modified in the light of new evidence, then the postulate; constructed to explain it was modified.

In other cases, however, debate focused on which of two postulates to modify. Castle, for example, accepted Postulate 2 that a tight connection existed between the unit and character and argued that variability of character implied variability of the gene, thereby denying Postulate 5. Muller countered by retaining Postulate 5 and denying a strict version of Postulate 2 by postulating effects of modifying genes on a character. This was a legitimate debate about where the locus of modification should be and both modifications had arguments in their favor. Appeals were made to *ad hocness*, simplicity, and consistency with other results. A crucial experimental test ruled in favor of Muller's alternative.

When Cuénot found an exception to the 3:1 ratios, there was agreement as to which set of postulates needed modifying: Postulates 5–8 which had been constructed to account for the 3:1 ratios. But debate occurred as to which should be modified. The modifications ranged from a very fundamental change to leaving the postulates intact and explaining the case away as an unusual one. Fundamentality here is analyzed as that which would require the most extensive modification of other postulates.² If, for example, Morgan's denial of purity of the gametes had prevailed, it would have had consequences for the postulate claiming that genes are not modified by being together in the hybrid (Postulate 5). However, Cuénot's proposal of selective fertilization (modification of Postulate 8) left Postulate 5 intact. Castle's explanation of the exceptions to 3:1 ratios left all the postulates intact and explained the particular case as due to an exceptional occurrence. Deviations from the 3:1 ratios were not common, lending credence to Castle's less fundamental change.

² Dudley Shapere in a recent paper discussed another case (the problem of solar neutrinos) and made a penetrating comment relevant to this question of fundamentality:

If the predictions of our hypotheses disagree with our observations (as they have in the case of solar neutrinos), it is not necessarily the hypothesis under examination (in that case, hypotheses about the processes of stellar energy production) which is at fault; it may be any portion of the theoretical and instrumental ingredients which form the background of the experiment. The point is, of course, not new: but what needs to be made clear and precise is that there exists a rough rationale for the order in which we subject the ingredient accepted ideas to suspicion, and correlatively, an order in which we propose new alternatives. We begin by suspecting those ingredients which are most likely to be a fault, and least costly to give up; if the difficulty persists, the threat penetrates deeper and deeper into the structure of accepted beliefs involved in the purported observation and background. (SHAPERE, 1980, p. 79).

Conclusion

I have examined one case from the history of biology in detail. What can be concluded on the basis of only one case? I cannot prove that I have found general patterns of reasoning in science. Nor do I have a basis for prescribing how successful science is to be done, a worthy, but perhaps unattainable, goal of philosophy of science. But this case does belie the general claim that one never has any reasons for believing that one rather than another postulate of a theory is in need of modification in the light of new evidence. Furthermore, the reasoning patterns found in this case serve as examples which can be looked for in other cases to test their generality and which suggest strategies for achieving certain goals in theory construction and modification.

By viewing theory construction as a modular process, one can see particular postulates are related to particular domain items. If the item changes, then the first locus to be considered for possible modification is the postulate or postulates which account for that item. The case examined shows that scientists usually agree on one or a few postulates as those in need of modification. The task then becomes to devise and choose among alternative ways of modifying that localized postulate. Devising alternatives is the process of discovery; choosing among them involves various aspects of theory choice or confirmation. Both are intimately connected in producing a modified theory. Although no algorithm is known for producing correct modifications, we can imagine a scientist (more likely a computer program³ these days) who employs a systematic strategy when faced with exceptions.

Several goals are chosen. First, the theory must account for the domain items; exceptions or additions to the domain may thus necessitate changes in the theory. Secondly, a consistent set of postulates must result from any modification. Thirdly, an exhaustive list of alternative modifications should be devised. In actual practice it is often hard to find one or two, much less produce an exhaustive set. Usually one scientist devises only one alternative and argues for it against competitors proposed by another scientist. However, more expansive minds devise their own alternative hypotheses (e.g., see a recent article by Francis CRICK (1979) on the possible mechanisms for eliminating intervening sequences from genes

³ Work along these lines is being done by Bruce Buchanan of the Computer Science Department at Stanford University, for example in his recent paper (BUCHANAN, 1978).

prior to their expression). For pragmatic reasons (or maybe something stronger, such as, we can only gain new knowledge when we have a base of older, better established knowledge) an economy of effort is desirable: make the least fundamental modifications first, then proceed to more fundamental ones.

With these goals and directives, we can devise a systematic strategy for modifying a theory to account for an exception to a domain item. Such a strategy is given in Table 3.

Table 3

| <i>Strategy for theory modification</i> |
|--|
| These steps should be taken in the order given. When an exception to a theory arises: |
| (i) Confirm the experimental results to be sure it is an exception. |
| (ii) See if such an exception arises only in the system studied (e. g., one character in one species) or whether it is found in other systems (e. g., other characters, other species). |
| (iii) Locate the postulate constructed to account for the domain item to which it is an exception. See if the postulate can be "complicated" to explain the exception (e. g., add another variable). |
| (iv) If (iii) failed to locate only one postulate, then examine the two or more postulates involved. Devise an exhaustive list of modifications possible, making sure that a consistent set of postulates results from each modification. Choose the modification that has least effect on the other postulates and test it experimentally. If it fails, then make the next least fundamental modification, test, and so on. |
| (v) If a modification cannot be used to make new predictions by which it can be tested, because it only accounts for the exception for which it was devised, then it is to be shelved as an unacceptable <i>ad hoc</i> modification. |

In summary, this paper has examined questions about reasoning in scientific change. More specifically, it has focused on theory construction and modification and argued that in some cases, at least, rational procedures exist for localizing parts of a theory in need of modification in the light of exceptions. Furthermore, it has suggested a strategy for carrying out those modifications, given certain goals of science, such as accounting for the data and devising consistent theories.

Acknowledgment

This research was supported by the History and Philosophy of Science Program of the U.S. National Science Foundation (Grant Soc 77-23476).

References

- BATESON, W., 1902, *Mendel's principles of heredity—A defense* (University Press, Cambridge, England)
- BATESON, W., 1909, *Mendel's principles of heredity* (University Press, Cambridge, England)
- BUCHANAN, B., 1978, *Steps toward mechanizing discovery*, Proceedings from a Conference on Logic of Discovery and Diagnosis in Medicine, Pittsburgh, October 1978 (Stanford Heuristic Programming Project Memo HPP-79-28)
- CASTLE, W. E., and C. C. LITTLE, 1910, *On a modified Mendelian ratio among yellow mice*, *Science*, vol. 32, pp. 868–870
- CASTLE, W. E., and J. C. PHILLIPS, 1914, *Piebald rats and selection: An experimental test of the effectiveness of selection and of the theory of gametic purity in Mendelian crosses*, Carnegie Institute of Washington Publication, No. 196, pp. 51–55
- CASTLE, W. E., 1916, *Pure lines and selection*, *Journal of Heredity*, vol. 5, pp. 93–97
- CASTLE, W. E., 1919, *Piebald rats and the theory of genes*, *Proceedings of the National Academy of Sciences*, vol. 5, pp. 126–130
- CASTLE, W. E., 1951, *The beginnings of Mendelism in America*, in: *Genetics in the 20th Century*, ed. Leslie C. Dunn (New York, Macmillan), pp. 59–76
- CRICK, F., 1979, *Split genes and RNA splicing*, *Science*, vol. 204, pp. 264–271
- CUÉNOT, L., 1905, *Les races pures et leurs combinaisons chez les souris*, *Archives de Zoologie Experiméntale et Générale*, 4 Série, T. 111, pp. CXXIII–CXXXII
- DARDEN, L., 1974, *Reasoning in scientific change: The field of genetics at its beginnings*, Ph. D. Dissertation. (The University of Chicago, Chicago, Illinois)
- DARDEN, L., 1976, *Reasoning in scientific change: Charles Darwin, Hugo de Vries, and the discovery of segregation*, *Studies in the History and Philosophy of Science*, vol. 7, pp. 127–169
- DARDEN, L., 1977, *William Bateson and the promise of Mendelism*, *Journal of the History of Biology*, vol. 10, pp. 87–106
- DARDEN, L., and N. MAULL, 1977, *Interfield theories*, *Philosophy of Science*, vol. 44, pp. 43–64
- DARDEN, L., 1978, *Discoveries and the emergence of new fields in science*, PSA 1978, vol. 1, Philosophy of Science Association (East Lansing, Michigan), pp. 149–160
- DARDEN, L., *Theory construction in genetics*, in: *Scientific Discovery Case Studies*, ed. T. Nickles (Reidel Dordrecht) (forthcoming)
- DARWIN, C., 1868, *Provisional hypothesis of pangenesis*, in: *The Variation of Animals and Plants Under Domestication*, vol. 2, ch. 27 (Orange Judd and Co., New York)
- DUHEM, P., 1914, *The aim and structure of physical theory*, trans. F. P. Wiener, 1962 (Atheneum, New York)
- EAST, E., 1910, *A Mendelian interpretation of variation that is apparently continuous*, *American Naturalist*, vol. 44, pp. 65–82
- JOHANNSEN, W., 1903, *Heredity in populations and pure lines*, selections trans. and reprinted in: *Classic Papers in Genetics*, ed. James A. Peters (Prentice Hall, Engelwood Cliffs, N. J., 1959), pp. 20–26
- JOHANNSEN, W., 1909, *Elemente der Exakten Erblichkeitslehre* (G. Fischer, Jena)
- LAKATOS, I., 1970, *Falsification and the methodology of scientific research programmes*, in: *Criticism and the Growth of Knowledge*, ed. I. Lakatos and Alan Musgrave (The University Press, Cambridge, England), pp. 91–195

- MENDEL, G., 1866, *Experiments on plant hybrids*, reprinted in: The Origin of Genetics, A Mendel Source Book, eds. Curt Stern and Eva Sherwood (W. H. Freeman and Company, San Francisco), pp. 1–48
- MORGAN, T. H., 1905, *The assumed purity of the germ cells in Mendelian results*, Science vol. 22, pp. 877–879
- MORGAN, T. H., 1909, *Recent experiments on the inheritance of coat colors in mice*, American Naturalist, vol. 43, pp. 494–510
- MORGAN, T. H., 1911, *The influence of heredity and of environment in determining the coat colors in mice*, New York Academy of Science Annals, vol. 21, pp. 87–117
- MORGAN, T. H., 1914, *Multiple allelomorphs in mice*, American Naturalist, vol. 48, pp. 449–58
- MORGAN, T. H., 1926, *The theory of the gene* (Yale University Press, New Haven)
- MULLER, H. J., 1914, *The bearing of the selection experiments of Castle and Phillips on the variability of genes*, American Naturalist 48, reprinted in: MULLER, 1962, pp. 61–69
- MULLER, H. J., 1962, *Studies in genetics, The selected papers of H. J. Muller* (Indiana University Press, Bloomington)
- QUINN, P., 1974, *What Duhem really meant*, in: Methodological and Historical Essays in the Natural and Social Sciences, Proceedings of the Boston Colloquium for the Philosophy of Science 1969–1972, eds. R. S. Cohen and M. Wartofsky, vol. XIV (Reidel, Dordrecht), pp. 33–56
- SHAPERE, D., 1974, *Scientific theories and their domains*, in: The Structure of Scientific Theories, ed. Frederick Suppe (University of Illinois Press, Urbana), pp. 518–565
- SHAPERE, D., 1980, *The character of scientific change*, Scientific Discovery, Logic, and Rationality, pp. 61–101, ed. T. Nickles (Reidel, Dordrecht)
- TOULMIN, S., 1972, *Human understanding*, vol. I (Princeton University Press, Princeton)

THE REALITY OF BIOLOGICAL SPECIES: A CONCEPTUALISTIC AND A SYSTEMIC APPROACH

OSVALDO A. REIG

Division of Biological Sciences, Simón Bolívar University, Caracas, Venezuela

1. Introduction

Species are one of the basic objects of investigation in evolutionary biology. Therefore, it would seem that postulating that species exist is a necessary point of departure for any science of evolution to make sense. Hence, it is not surprising that most contemporary biologists, when asked about what species are, endorse without qualm the view that they are real entities, whatever the meaning they ascribe to these words. And in some cases, particularly when the question is addressed to scientists enmeshed in the nets of logical empiricism, the reply is that it is a useless question or, at most, a subject only worthy of discussion in conclaves of excentric philosophers.

Fortunately, the holders of the latter contemptuous view are not many nowadays. For if it were a common belief among biologists, it might have produced some kind of bad reputation for these scientists in learned circles. No educated scholar would doubt that the question about the kind of existence species have is a genuine metabiological problem. Indeed, it was a subject of profound arguing among the most influential Western thinkers, from Plato and Aristotle to our days. Certainly, the ontological question involved is not a simple one, as it refers to some of the more complex queries of philosophy: the kinds of being and existence, the problem of universals, the nature of entities, if any, of the external world, the relationships between knowing and what is known, and so on.

To produce a new piece of written discussion in such a much debated subject may evoke a feeling of suspicion in a demanding and well-informed reader. It may comfort him to know that I also feel disturbed by the scope of my enterprise, and that I only dared to submit the forthcoming re-

flexions to a learned audience after convincing myself that biologists have good reasons to enter into philosophical disputes about biological problems, because they have first hand experience in the subject-matters involved in such discussions. Only secondarily, I hope to bring a tentative solution to the arduous problem of the reality of species, convinced as I am that taxonomic species are better understood under a conceptualist epistemology, and that they refer to natural entities of the external world whose attributes can be elucidated under a systemic ontology.

2. Statement of the problem and brief retrospect

Our aim is to explore the ontological nature of biological species. The subject implies an enquiry about the kind of objects to which the name species applies, and about the kind of existence of these objects. Are species endowed with real existence, or are they artificial constructs of the mind? If they are real entities, in what sense are they real? Is the name species applied to a single sort of thing in biology, or is it a term which denominates different kinds of biological objects?

The metaphysical and logical concept of species underlies the concept of biological species and, as such, it has a very long history, being an important topic in the philosophies of Plato and Aristotle. Starting from this tradition, during many centuries of western thought the nature of species has been explained under the tenets of several forms of idealistic realism, with their Platonic, Aristotelian or Scholastic variants. Under this conception, species were real in the sense that they were essences or forms existing in the world of ideas (HULL, 1965). This sort of realism prevailed during the XIII Century in Christian thinking, under the influence of Duns Scotus and St. Thomas.

Connected with the idea that species were transcendent essences in the realm of ideas, was the common belief of many centuries, that the species we perceive in the external world are deceitful appearances, ephemeral and erratic (ZIRKLE, 1959, p. 638). Certainly, this conception backed the belief in spontaneous generation and the transmutability of biological species.

It took a long time to eradicate those views. The gradual outcome of modern science in the XVI and XVII centuries, with its reaction against Aristotelianism and the claim that observation and experiment were the grounds on which the knowledge of nature should be based, opened the road to another way of looking at biological species. The early modern

biologists were not comfortable with the idea that the species they studied in nature were mere deceitful shadows of transcendent essences unattainable through scientific procedures, and they started to look at their relationships with the material world. By the middle of the XVII century Swammerdam strongly opposed the old idea of spontaneous generation, and, a little later, the experiments of Redi contributed to the rebuttal of that view. In the same century, John Ray established belief in the rigorous delimitation of species as discrete entities knowable by objective attributes, thus affording grounds for abandoning the idea of transmutability.

However, we must arrive at the XVIII century to find the first expression of the two main confronting modern views about the nature of biological species: nominalism and what can be called "natural" realism.

Probably the main conscious reaction against idealistic realism was nominalism or terminism, in the form of a belated assimilation of Ockam's empirical views (HULL, 1967). In the XIV century, William of Ockam advocated that the objects of real science were individuals, claiming that the reality is a reality of individuals, and that there were no entities whatsoever between terms and the real individuals to which they refer. Applied to biological species nominalism holds that they are nothing but names applied to sets of individual organisms, and that only those individual organisms existed in reality. Buffon was probably the earliest explicit exponent of nominalism as regards biological species (HULL, 1967). However, he was not quite a consistent thinker, and we can find in his writings many statements in contradiction with that view. In the second half of the century and later, at the beginning of the XIX century, nominalism is clearly stated by Robinet and Lamarck. The first wrote, in 1768: "Il n'existe pas que l'individu. L'espèce des naturalistes n'est qu'une illusion" (ROBINET, 1768). Lamarck's view on the nature of species is well known. It suffices here to recall his famous statement in his opening address of 1806: "souvenons-nous que rien de tout cela n'est dans la nature; qu'elle ne connaît ni classe, ni ordres, ni genres, ni espèces..., il n'y a réellement que des individus et des races diverses qui se nuancent dans tous les degrés de l'organisation" (LAMARCK, 1907).

We can trace the opposed realistic view of species in Robinet's contemporary, Carl Linnaeus. Linnaeus established the concept of species as constant entities provided with objective attributes, and sharply delimited. Linnaeus was not a metaphysical thinker, and his realistic conception of species was deprived of the invocation of transcendent essences,

notwithstanding his use of Aristotelian logic. It is therefore legitimate to maintain that, by claiming that species are discrete and constant entities, he was espousing a kind of empirical or natural realism, fundamentally similar to the position advocated by Guillaume de Champeaux in the XII Century (FERRATER MORA, 1965).

Thus, since the middle of the XVIII century, more than two hundred years ago, the problem of the nature of biological species has been tossed back and forth between nominalists and different sorts of realists in what sometimes appeared to be an impossible dialogue between people who were both opinionated and deaf. The details of this discussion have been well treated elsewhere (see MAYR, 1957; HULL, 1965, 1967). It is of interest here only to recall that it has not been devoid of some paradoxes. Thus, Darwin himself, one of the first seriously to attack the problem of the origin of species, espoused at the same time the idea that species were purely arbitrary and subjective (DARWIN, 1859, pp. 484-85 and elsewhere). Similarly, it is striking to realize that species were divested of reality by such representative early Mendelians as Bessey, Bateson and de Vries (BESSEY, 1908; MAYR, 1957), since in fact it was the marriage between darwinism and mendelism what fixed the starting point of the present conception of species as objective evolutionary entities.

Ernst Mayr, in his introductory chapter to the now classic volume *The species problem* (MAYR, op. cit.) was surprised by the lack of solution to a question to which such an immense amount of time and thought has been devoted. He spoke of a "hidden reason" behind such continuity of disagreements, and hinted that the reason lay in the fact that there was more than one kind of species involved in the discussion. He dedicated himself to elucidate the various meanings of the name "species", advancing there his well-known distinction between typological, non-dimensional and polytypic or "biological" species concepts (see also MAYR, 1963, 1969a).

I think, however, that Mayr had no reasons to be struck by the lasting continuance of waverings on the nature of species. It was an expression of the classic problem of universals, which is not a simple problem and which, after many centuries of discussion is still a genuine and unsettled metaphysical problem. Moreover, I believe that the hidden reason for, such long-lasting disagreements may lie in the fact that in such discussions the issue was confused by a misapprehension of a more subtle, but far more important difference in the usage of the word "species". This sort of difference was already noticed by authors who distinguished between

the application of the term species to units of identification in the field of biological taxonomy, and to units of evolution in the field of evolutionary theory (BLACKWELDER, 1962; DOBZHANSKY, 1951; GILMOUR, 1961; GRANT, 1963; SONNEBORN, 1957). But, I also believe that this distinction has still not been thoroughly analyzed, and that by performing this analysis important consequences can be drawn out.

A point of departure for such an analysis is to consider the distinction between taxonomic species (the species as units of identification) and evolutionary, "biological" or Mendelian species (SMITH, 1958) (which, adopting a term proposed by Cain, 1954, I shall call here for brevity's sake "biospecies"), in the framework of the distinction between a cognitive realm and an ontic realm. As a consequence of the counter-revolution against logical empiricism, it is now widely admitted that metaphysical assumptions play an important role in sciences (BUNGE, 1971, 1974; WATKINS, 1975). One of the more important metaphysical presuppositions that guides scientific research is the ontological assertion that there is an external world made of natural things, entities which have properties of their own and which are submitted to processes of change. Science approaches the knowledge of those entities and processes of the external world—the ontic realm—through scientific concepts, models and theories. The latter belong to the cognitive or conceptual realm, which represents a perfectible picture of the ontic realm, the description and understanding of which is the aim of any scientific endeavour.

3. Some fundamental concepts of biological taxonomy

The logical operation of classifying is one of the means for a conceptual approach to understanding the diversity of natural things. That portion of the external world which is organized as living matter, exhibits an overwhelming diversity of natural entities. V. GRANT (1963, p. 82) calculated that our planet is inhabited by about 4.5 millions different kinds of animals and plants. Thus, the science of biological classification—the biological taxonomy or systematic biology—grew up as an important part of biological sciences, developing its own set of formal and conceptual devices.

The Linnean hierarchy is the formal framework by reference to which organic diversity is classified. This hierarchy is made up of an orderly set of conceptual items, the taxonomic categories. They follow a definite order of subordination, and are connected among them by the logical

relation of inclusion. They are the well-known categories of Kingdom, Phylum, Class, Order, Family, Genus and Species, disregarding several intermediate ones which are used sometimes in requirement of more complicated patterns of diversity. Each of these categories is a class concept, the members of which are other class concepts, the taxa, which we shall treat in the next paragraph. But they are peculiar concepts, as they have extension, but are not defined in regard to a set of properties, but just in regard to the rank they have in the hierarchical system. Therefore, taxonomic categories are pure abstractions of the mind; they only exist in the conceptual realm, and they have no direct referents in the ontic realm.

Biological classifications are systems of taxa ordered in subordinated ranks in accordance with the ranks of the Linnean hierarchy, as any taxon is a member of a given taxonomic category. The taxa are class concepts which refer to groups of real organisms. Any taxon is distinguished by reference to its concrete representatives in nature. Thus, the Order Primates is a taxon, and its extension is the set of all monkeys, lemurs, apes and hominids, living or extinct, that exist or have existed in our planet. And the Class Arachnida is a taxon concept to the extension of which belong all the spiders, mites, scorpions, harversmen, sticks and the like which inhabits our world or which lived in the past. Besides, any taxon is defined by reference to the taxonomic category to which they belong. Thus, the Class Arachnida is a member of the taxonomic category Class, and the Order Primates is a member of the category Order, just as the species *Homo sapiens* belongs to the category Species. Thus I propose the following definition of taxon: "A taxon is a class concept in biological classifications which belongs to a given taxonomic category, and which refers to a given group of organisms". Other proposed definitions, as that of SIMPSON (1961, p. 19) or the one of MAYR (1969b, p. 4), even when they emphasize that taxa are related to groups of concrete organisms, identify the taxa with the groups of organisms, failing to recognize that taxa are concepts and not things, and are vitiated by circularity, as they confound the *definiens* with the *definiendum*. It is important here to notice two fundamental differences between taxa and taxonomic categories. The taxa are concepts which refer to groups of organisms existing in the realm of reality, whereas taxonomic categories are deprived of direct referents in nature. Additionally, the former, in contrast with the latter, have both extension and intension, as membership to a given taxon is determined by the possession of the taxon-member of a set of

properties which define the taxon-concept. Thus, for a given organism to be a member of the taxon Arachnida, it must have its body divided in prosoma and opisthosoma, with six somites in the prosoma and twelve somites in the opisthosoma, six pairs of articulated prosomal appendages, and so on. The following example may help towards a better understanding of the difference between taxa and taxonomic categories, which some people find difficult to grasp.

The taxon *Mus musculus* belongs to the taxonomic category Species, but it has as referents all the individual organisms that live or have lived in nature and which belong to the species of mice *Mus musculus*, because they satisfy the possession of a set of attributes which define that species. In the same way, the genus *Mus* is a taxon of the category Genus, and its extension is made of all the individual organisms which match the definition of the genus *Mus*, irrespectively of the fact that they may belong to different species (*Mus musculus*, *Mus poschiavinus*, etc.). *Mus* and *Rattus* are different taxa but both are members of the category Genus, as well as *Mus musculus* and *Mus poschiavinus* are different taxa irrespectively of the fact that both are members of the category Species. Thus, taxa are connected to taxonomic categories by a relation of membership, and the same kind of relation holds for individual organisms with respect to their taxa. But the membership relationships between taxa and taxonomic categories is an interconceptual relationship; therefore, the latter are concepts of concepts, and, as we are dealing with class concepts, they are classes of classes. Contrarily, the relationships between taxa and the organisms which belong to them, are relationships between concepts and things. But we must also realize that taxa, whatever the category to which they belong (the species *Mus musculus*, the genus *Mus*, the family Muridae, the order Rodentia, and so on) are constructs that only exist in the conceptual realm, that is to say, their existence is dependent of a cognizant agent. In this respect, they sharply differ from the organisms to which they refer, as the latter have a substantial existence: they are located in space, they are endowed with measurable energy, they can be seen, smelt, heard or touched, they are submitted to processes of change, and they are able to assemble in biological systems (BUNGE, 1976, p. 203).

However, the ontological question of the nature of the referents of taxa is not fully answered by assessing the properties that give substantial existence to individual organisms. It can be accepted that individual organisms are material things existing at the ontic realm independently of any act of knowledge, without accepting that their taxonomic groupings

correspond with natural groups. As a matter of fact, empirical nominalism would admit that individuals do exist in the real world, but it will also claim that the grouping of individuals into an orderly set of classes is an artificial construction which does not reflect any kind of group-entity in nature. Supporters of a realistic conception of taxa will argue that the taxonomic system is a reconstruction of a given pattern of level structure of the organic diversity in nature, and that that pattern is the product of an objective process occurring in the material world: organic evolution. This contention may sound like a reminiscence of the parallelism between the logic hierarchy and the ontologic hierarchy in Porphyrium's tree, but advocates of a parallel between the taxonomic hierarchy and the level structure of organic diversity are not necessarily fond of old essences kept in Aristotelian bottles. In fact, one of the points of departure of the theory of evolution was the phenomenological hypothesis explaining the classificability of organic diversity in a hierarchical system as a reflection of a pattern of multiple level relationships of affinity in organic diversity, resulting from a process of gradual diversification from a remote common ancestor (HANSON, 1961, p. 34).

If we take the realistic argument and the evolutionary explanation for granted, it would seem that we should accept that all taxa are real, or better, following our conceptualistic position, that they are natural classes. I believe, however, that it is possible to demonstrate that taxa of supraspecific rank are of a different kind from taxa of species rank, as regards the naturalness of their referents. In fact, the only valid reason to suppose that taxa of supraspecific rank are not artificial constructs, is that their members share with each other a more proximate common ancestor than any of them with members of different taxa of the same rank. But they could hardly be considered to refer to discrete and integrated assemblages of individual organisms held together by interconnecting relationships, as I shall try to demonstrate is the case for taxa of the category Species. Actually, the possession of a common ancestor is not necessarily determinant of the maintainance of cohesion between the descendants. The grandchildren of a couple of grandparents share with each other a more proximate ancestry than any of them share with the grandchildren of a different couple. However, it does not follow that because of that, they must keep being connected with each other in natural associations, like belonging to the same church, being members of the same corporation, or of the same political party, or of any other economical, ideological or social structure. In the same way, it does not necessarily happen that

different individual organisms belonging to the same genus, or the same family, are interconnected in the economy of nature as members of the same supraorganismic system of integration. In fact, species of different genera, and usually belonging to different families or to different higher-order taxa, are held together in the structure of a biological community, irrespective of their taxonomic relationships. There are ecological relationships that link the different species of a community in a tightly cohesive supraorganismal unit. But the community is an ecological unit, and not a taxonomic one.

Thus, apart from being related because they share more proximate common ancestor, the subordinated members a supraspecific taxon, are not necessarily connected with each other by objective links in the world where their representatives live. Supraspecific taxa cannot, therefore, be considered as having as referents natural entities of the external world. Contrarily, the taxa of species rank reflect assemblages of individual organisms which share with each other a definable set of intraspecific relationships. These relationships make the assemblage behave in nature as a cohesive supraorganismal entity of its own: an integrated biological system having properties which are not found in its individual components, and which emerge from the intraspecific relationships (see below, Section 5).

Therefore, we can conclude that it seems at first analysis that taxa of supraspecific level and taxa of species rank refer to objects of a different kind. A certain level of nominalism seems to be attributable to the former (BUNGE, 1967, p. 79), but not to the latter. I am convinced, however, that the ontological nature of higher taxa deserves further analysis. Unfortunately, a deeper scrutiny of this problem is out of our present aims, and we must switch to further discussing the topic of the nature of taxa of species rank.

4. The epistemic link between taxonomic species and biospecies

It follows from what we have so far discussed, that the word "species" is applied to three different kinds of objects:

- (1) One of the taxonomic categories in the Linnean hierarchy;
- (2) Taxonomic species, i.e., conceptual constructs having empirical referents of the kind of natural assemblages of individual organisms tightly connected with biospecies; and
- (3) those assemblages in the real world, the biospecies.

We have also seen that the nature of taxonomic categories is that of

a logical class, formal and abstract, and that the taxa are conceptual constructs which refer to concrete thinks, either arbitrary or semi-arbitrary sets of individual organisms, or integrated supraorganismal entities. Obviously, the Species as a category, and the taxa of species rank only exist in the conceptual realm: their existence is dependent on the thinking activity of a cognizant agent. There is no question, therefore, that they do not exist at the ontic realm. We have also advanced that we believe that species-taxa are not artificial classificatory concepts, but that they are natural classes, in the sense that they reflect collective biological entities that have the properties of material systems of the external world. The latter are the kind of objects that several authors have in mind when they speak of species as evolutionary units.

It has been repeatedly suggested (BURMA, 1954; BLACKWELDER, 1962; SONNEBORN, 1957) that the species concept of taxonomists (our taxonomical species) is quite different from the species concept as used by evolutionary biologists (our biospecies). The former would be groups made up from individual organisms sharing like attributes, usually morphological attributes, whereas the latter would refer to natural evolving populations. BLACKWELDER (1962, p. 34) claimed that misunderstanding and confusion are inevitable, as we have two different things represented by a single term. SONNEBORN (1957) went even farther, and he directly proposed to call by another name, "syngen", the species as an evolutionary unit.

In my belief, this way of posing the question leads to a still greater misunderstanding. It is true that taxonomists most frequently use the concept of species to refer to groups of organisms sharing similar morphological characters, and that they only rarely take into account the evolutionary and populational attributes of their species. But the property of being alike is a quality which results from populational and evolutionary links among the organisms which are alike. It is the expression, at the phenotypic level, of a common gene pool which maintains its homogeneity within a population through the action of a given set of evolutionary factors. Therefore, when a taxonomist assorts individual organisms into species-taxa on the basis of their shared morphological attributes, he is also distinguishing throughout their affinity, sets of individual organisms which are tied together in nature by genetic, populational and evolutionary relationships.

It is also true that taxonomical species may not completely correspond to evolutionary species, and that in many cases morphological likeness

or difference may be misleading in regard to the differentiation of biospecies. Several biospecies exhibit morphological polymorphisms, and different morphs of the same biospecies may be at least as different in morphology as members of different biospecies. Moreover, cases of sibling or cryptic species are continuously reported: they are biospecies which attained complete reproductive isolation, but are almost impossible to differentiate in their morphology. But these shortcomings of the morphological method are only a reflection of a well-established conclusion among biologists: morphological differentiation even when it is a common consequence of the genetic differentiation is not necessarily related to the latter by equivalent amounts of expression. In this respect, the operation of sorting out individual organism into different taxonomical species on the basis of their degree of morphological affinity, may be considered as a legitimate, but perfectible, way of approaching the recognition of biospecies in nature. The accuracy of this recognition may further improve when a larger set of biospecies attributes are considered (REIG, 1968).

That this is what actually happens in the taxonomical practice, is demonstrated by the fluidity of taxonomical species. Every taxonomist will agree that the latter are not stable conceptual constructs: they change in extension and in intension as a consequence of the improvement of our knowledge. New species are currently proposed by zoologists and botanists for organisms that had previously been placed in other species, or which were included under a broader species concept. It also happens frequently that organisms that were classified as different taxonomical species are later lumped into a single species concept after reaching a deeper knowledge of their properties. Diagnoses of taxonomical species—which is equivalent to the assessment of their intension as taxon-concepts—are frequently modified to include new attributes or to withdraw irrelevant features.

All this fluidity of taxonomic species would be highly inconvenient, and even senseless, if the species of taxonomists should be considered to be artificial constructs. In fact, the changing nature of species at the conceptual level must be taken as a clear indication that taxonomists aim to deal with taxonomic species which match as well as possible to species as natural assemblages of evolving organisms. Therefore, we must conclude that the species of taxonomists are different as regards the species of evolutionary biologists only in the sense that the former are a conceptual and perfectible reconstruction of the latter.

We may now restate the problem by asserting that taxonomic species

and biospecies are connected by an epistemic link. The former change with the improvement of biological systematics, tending to become an increasingly closer reconstruction of the latter as available information and theory improves. Therefore, the statement that a given set of individual organisms or a given set of population samples belong to a taxon of species rank, may be considered as the statement of a hypothesis about the ontological status of those collections of individuals or population samples. The hypothesis may be based only on the observation of common traits shared by the individuals of the samples. To test the hypothesis, the evolutionary taxonomist must investigate if, as is often the case, the sharing of like features is correlated with other critical attributes of biospecies, which will hinge, as we shall see below, on the possibility of inferring the occurrence of reproductive isolation. If evidence of such other attributes is found, the hypothesis becomes corroborated. Otherwise, it is rejected, which, in common practice, means to propose new arrangements for the same organisms or samples.

It is now necessary to explore a little deeper what are the attributes of biospecies which justify my claim that they are natural entities of the external world by reference to which taxonomic species are constructed.

5. The nature of biospecies as biosystems

As already noticed, taxonomic nominalism maintains that in nature only individual organisms exist, and therefore, that nothing of the kind of species can be recognized in the ontic realm. This contention implies a very restricted idea of physical individuality, and it fails to recognize the existence of levels or grades of individuation (FERRATER MORA, 1962). In fact, individual organisms represent just one level of individuality, and they also can be thought of as assemblages of individual cells, in the same way as cells may be considered as assemblages of individual macromolecules, just as the latter are made of interconnected individual atoms. Organisms, cells, macromolecules and atoms each represent a given level of individuality, and they are, in their own level, particular entities with the characteristics of physical systems: integrated and cohesive units of different level of complexity.

But in the same manner as organisms, cells, macromolecules and atoms are individual entities in increasingly simpler levels of complexity, it can also be realized that there are individual entities in supraorganismal levels

of integration: populations, biospecies, communities, ecosystems, the biosphere as a whole, the solar system, the Milky Way, and so on.

It is precisely my contention that biospecies are material entities in one of those supraorganismal levels of integration, and that each biospecies possesses substantial existence as an individual biosystem of organisms. A given biospecies is composed, of course, of individual organisms which are its immediate components, in the same way that the immediate components of a molecule are its atoms. But as a molecule is more than an aggregate of atoms, the individual organisms which belong to a biospecies are more than a mere aggregate of isolated parts. Contrarily, they are held together by interrelationships which act as cohesive bonds, in such a way that the assemblage has properties of its own emerging from the interrelationships of the individual components, therefore it behaves as an integrated whole.

In the present inquiry, I imply by emergence and wholeness concepts which have little in common with Lloyd MORGAN's (1923) emergent evolution, or MEYER-ABICH's (1948) holism. Emergent properties of systems have nothing mysterious: they are able to be known and analysed as properties resulting from the interactions of the immediate components of the system. In spite of the fact that they may not be characteristics of those components, they are rooted in them and therefore they are liable to scientific scrutiny. Nor are integrated wholes the result of special totalizing forces only knowable by means of intuition. Under the tenets of the systemic approach I adopt here, they represent different levels of structuring of matter having global properties which are knowable through analysis and synthesis (BUNGE, 1977).

Certainly I cannot pretend that my contention that biospecies are material systems is quite a new idea. Forty-eight years ago, VAVILOV (1931) advanced a very similar view. More recently, other Soviet authors defended the same idea under the inspiration of Bertalanffy's General Systems Theory (see, for instance, IANKOVSKY, 1966; SETROV, 1966; ZAVADSKY, 1966). Moreover, there has been also a tendency lately to admit that the concept of individuality can be extended to higher levels of organization, and to treat biological species as individuals (GHISELIN, 1975; HULL, 1974, 1976; SMART, 1963; VAN VALEN, 1976). Additionally, several authors have stressed that the reality of species hinges on its consideration as a gene-pool of interbreeding populations (HULL, 1977; LEHMAN, 1967; RUSE, 1969), and this concept underlies my contention that biospecies are biosystems. I believe, however, that a full-fledged

treatment of biospecies in the framework of the systemic approach (BUNGE, 1977) has not yet been attempted. And it seems to me that to look at biospecies from such an approach is not only a fruitful way to gain a deeper insight into their properties as interbreeding populations, but that it is also the most convincing way to demonstrate that they are material entities of nature.

It is time now to look at the properties and interrelationships of biospecies as individual biosystems. Unfortunately, to undertake such analysis in detail is outside the scope of the present paper, as it would require a switch to arguments full of biological technicalities. However, it is unavoidable to set forth a brief outline of the most important of those properties and interrelationships.

The most important factor that maintains the individuality of biospecies is reproductive isolation. The different mechanisms of reproductive isolation (see a modern survey in DOBZHANSKY *et al.*, 1977, pp. 170–179) have the effect of preserving the integrity of a distinct species-specific gene pool. This isolated and distinctive gene pool is the ultimate determinant of the observed character discontinuity of biospecies, and of their ecological and evolutionary properties. The observed characters can be morphological, physiological or behavioural, but whatever the case they are always rooted to genetic traits. From the perspective of population genetics, it is legitimate to visualize biospecies as collections of genes which are distributed from generation to generation among different individual organisms. The latter share an important relationship with each other: they are able to reproduce among them—the interbreeding concept—thus maintaining the continuity of the biospecies' gene pool, and allowing the assortment of different gene combinations in the new individuals which arise through sexual reproduction. This continual production of genetically different individual organisms, combined with natural selection, is a main factor in keeping biospecies adapted to their environments, and also in maintaining their adaptability to cope with the ongoing changes of the environment, which is the biospecies' insurance against extinction.

Now, the different mechanisms of reproductive isolation, the properties of the gene pool, in particular its capacity to generate intraspecific variability, and the interbreeding property, are all emergent properties of biospecies as integrated individual systems. This qualification comes from the fact that even when the mechanisms and properties are rooted in the

component individual organisms of each biospecies, they are not characteristic of them as individual components, but they become apparent only as a result of their interrelationships. Thus an individual organism is not endowed by itself with reproductive isolation towards an individual belonging to another biospecies: the property of being isolated only exists as a consequence of a complex of evolutionary forces which acted on its biospecies, and only makes sense in the framework of its relationships with other individuals. In the same way, an individual organism may be well adapted to its environment, but the continuity of the biospecies to which it belongs' ability of being adapted is not dependent on it as an individual, but on the total amount of the genetic characteristics of its biospecies. Within this biospecies, it merely represents a very limited fragment, and a temporal one of the biospecies' overall genetic potentiality.

Besides their population-genetical properties, though, the biospecies have ecological properties. Some of them are not emergent ones, as they result from the aggregation of the properties of the individual organisms. I would quote here the exploitation of a given portion of the environment—the species attribute of niche specificity—and the interspecific relationships of members of each biospecies within the framework of the community structure. Even when these are also global properties of each biospecies, they are also characteristic of any of their individual components as individual organisms. Other ecological properties of biospecies, as their bionomic strategies (SOUTHWOOD, 1976), the intrinsic rate of natural increase (BIRCH, 1948) and the other life-table parameters, are to be considered as emergent properties. It is not an isolated individual which may be considered as a K or a r -strategist, but the biospecies population to which it belongs as an integral entity. In the same way, the intrinsic rate of natural increase is not a property of any of the individual organisms, but rather an emergent quality of the biospecies' populations as integrated wholes, even though this rate is dependent on the net reproductive rate (R_0) and the generation time (T_0), which are dependent on the average rates of births and deaths of individual organisms.

Thus, biospecies are to be considered as biosystems, because, even when they have some properties which result from the aggregation of individual properties of their component organisms, they have also characteristic emergent properties which ultimately derive from the relationships among the latter (BUNGE, 1979, p. 14). Reproductive isolation, interbreeding, the possession of a common gene pool, adaptedness, biodemo-

graphic parameters and bionomic strategies are all emergent properties of biospecies which, together with niche specificity and community relationships, make them integrated biosystems, and not sheer aggregates of individual organisms.

6. Self-defense

In discussing these views among some of my colleagues, I received several enriching comments, but also some significant objections. Among the latter, there are two which I consider worthy of attempting a rebuttal. I shall call them the *anti-transference argument* and the *gradualistic argument*.

The anti-transference argument claims that in treating biospecies as biosystems, I am illegitimately transferring to the biospecies characteristics which belong to the populations; that the populations are actually biosystems, whereas species are nothing more than conventional sets of populations built up for the sake of classifying. At first sight, it may seem that there are cogent reasons in favour of this arguments. One speaks of population genetics and population ecology, but not of species genetics and species ecology, and these names reflect the widely accepted idea that populations are the actual units of first level supraorganismal integration. However, there is here a matter of tradition more than a point of fact, as every evolutionary biologist will be ready to accept that populations are integral parts of biospecies.

I am ready to admit that species as taxa are frequently constructed on the basis of sets of population samples. But it is also evident to me that when we classify a given set of populations as a taxon of species rank, we are postulating that those populations are held together by the possession of a common gene pool, maintained by a certain amount of actual or potential gene flow, and, therefore, that any member of a population of a given biospecies is able to reproduce with any member of the other populations belonging to the same biospecies. Thus, populations are integrated into biospecies by objective relationships. Actually populations must be considered as spatially and temporally differentiated component subunits of biospecies. As such, they may behave as relatively unstable subunits, being able to merge into each other in distribution or to separate geographically from each other depending on the dynamics of the changing environment. Sometimes, the spatial splitting is more lasting, and in those cases, it may elicit different degrees of raciation or subspeciation. But

only when a separate population differentiated so much that it attains reproductive isolation, must it be considered as making a distinct biosystem of its own, i.e., a distinct biospecies. In the absence of these mechanisms of isolation, populations are geographically—and sometimes only temporally—differentiated parts of one and the same biospecies.

The gradualistic argument claims that the evolutionary fluidity of biospecies is such that it becomes a matter of mere convention to fix limits to an allegedly distinct unit which gradually changes in time and eventually transform itself into a different unit. We are dealing with an argument of an old tradition. In biology, we can find an early expression of it in LAMARCK (1806). It is also an important piece of reasoning in BURMA's (1954) claim against the reality of biological species. The reader will also easily recognize that we are dealing here with the old metaphysical problem of the nature of becoming, which is rooted to the pre-socratic philosophers: the question whether becoming is to be understood as a process of accumulation of quantitative changes, or as a process of change of qualities.

In fact, it may be admitted that if biospecies would actually be steadily transformed into other biospecies through time without any implication of a qualitative change in the process, their delimitation would become a matter of sheer convention, whatever the attempts to make this operation as non-arbitrary and objective as possible (see IMBRIE, 1957, and the interesting discussion in HULL, 1965). However, this difficulty would not jeopardize the legitimacy of their attributes of discreteness and discontinuity in a given slice of evolutionary time, for instance the present. But the idea of the gradual merging of one biospecies into another through time, which was held by DARWIN (1859, p. 342) himself, and which seemed to be one of the corner stones of evolutionary theory, is now far from being a widely accepted view. The empirical support of this view, which is often called "phyletic gradualism" would be those recorded cases in palaeontology of the steady change of one morphologically defined biospecies into another through a continuous stratigraphic sequence. But most of those cases have been critically reexamined and reinterpreted otherwise (ELDREDGE, 1971; ELDREDGE & GOULD, 1972). As a consequence, it is now widely accepted that phyletic gradualism is, at most, an exceptional process. ELDREDGE & GOULD (see ELDREDGE, op. cit.; ELDREDGE & GOULD, op. cit., 1978) have recently set forth an alternative to phyletic gradualism which is in good keeping with population genetics and with the isolation theory of speciation. Following this model, the punctuated

equilibrium theory, biospecies are dynamic, but rather constant homeostatic units, which maintain their genetic and evolutionary characteristics during long periods of stasis, keeping themselves in qualitative invariance. New biospecies would arise by a sudden process of change taking place at the periphery of the distribution of the parental biospecies. Therefore, the speciation process would be an instance of becoming through a process of qualitative change. Thus, the origination of new biospecies would be equivalent to the emergence of an evolutionary novelty (REIG, 1977): new biospecies would arise as a consequence of a rapid evolutionary episode through which a new integrated and rather stable biosystem is created from a previous one. Hence, biospecies would be not only discontinuous in a given slice of evolutionary time; they would also be discontinuous through time.

7. Summary and concluding remarks

The main conclusions of this paper are the following:

- (1) The name "species" is applied in biology to three different kinds of objects, namely: (a) the Species, one of the taxonomic categories of the Linnean hierarchy; (b) the taxa of species rank, or taxonomic species, and, (c) supraorganismal evolutionary units, distinguished in this paper as biospecies.
- (2) (a) and (b) only exist as concepts: they belong to the realm of scientific concepts and theories and are not endowed with material existence in the ontic realm.
- (3) As with all taxonomic categories, the category Species is a class of classes; it is a classificatory concept the extension of which is made up of other classificatory concepts: the taxa of species rank; it is also defined by its position in the hierarchy, and not by intensional properties.
- (4) The taxonomic species, or taxa of species rank are also class concepts, but they are defined by intensional properties and their extension is not made up of other concepts, but of material things. They are conceptual constructs which reflect material entities existing in the ontic realm: groups of interconnected individual organisms, the biospecies.
- (5) Biospecies are more than sheer aggregates of individual organisms: they are cohesive and discrete evolutionary and ecological units with the attributes of supraorganismal individual entities.

- (6) Taxonomical species and biospecies are not entirely different entities. They are linked to each other by an epistemic bond, in the sense that the former are a perfectible reconstruction of the latter.
- (7) The ontological nature of biospecies is that of material biosystems, as they have systemic properties and are endowed with substantial existence. As in any system, biospecies show emergent properties arising from the interrelationships of their individual organismal components. But they also have properties resulting from the aggregation of the properties of their components.
- (8) Populations are temporary spatial subunits of biospecies; they are integrated to biospecies by actual or potential interbreeding which is a consequence of the lack of mechanisms of reproductive isolation.
- (9) Biospecies are homeostatic genetic-evolutionary biosystems, kept in dynamic equilibrium and qualitative invariance throughout their evolutionary existence. New biospecies originate by means of a sudden qualitative change determining the advent of isolating mechanisms and the resulting emergence of a new integrated biosystem.

Once it is accepted that taxonomic species are concepts referring to natural entities which behave in the world of nature as integrated biosystems of individual organisms, it would seem that the long-lasting controversy between nominalists and realists on the nature of biological species can be solved under the tenets of an epistemic conceptualism and a systemic ontology.

8. Acknowledgments

This paper is a side product of Project S1-0630, funded by the Venezuelan Council of Scientific Research (CONICIT). I am much obliged to Michel Ruse for giving me the opportunity to read this paper. I am also indebted to Mario Bunge for inspiration, to my wife, Estela Santilli and my friend Moritz Benado, for comments and criticism, and to Paul Berry for reading a first version of the manuscript.

References

- BESSEY, C. E., 1908, *The taxonomic aspect of the species question*, Amer. Nat., vol. 42, pp. 218-224
- BIRCH, L. C., 1948, *The intrinsic rate of natural increase of an insect population*, J. Anim. Ecol., vol. 17, pp. 15-26
- BLACWELDER, R. E., 1962, *Animal taxonomy and the new systematics*, Surv. Biol. Progr., vol. 4, pp. 1-57

- BUNGE, M. 1967, *Scientific research. I. The search for system* (Springer Verlag Inc., New York)
- BUNGE, M., 1971, *Is scientific metaphysics possible?*, J. Philos., vol. 68, pp. 507-520
- BUNGE, M., 1974, *Metaphysics and science*, General Syst., vol. 19, pp. 15-18
- BUNGE, M., 1976, *El ser no tiene sentido y el sentido no tiene ser: Notas para una conceptología*, Teorema, vol. 6, pp. 201-212
- BUNGE, M., 1977, *General systems and holism*, General Syst., vol. 22, pp. 87-90
- BUNGE, M., 1979, *A systems concept of society: beyond individualism and holism*, Theory and Decision, vol. 10, pp. 13-30
- BURMA, B. H., 1954, *Reality, existence and classification: a discussion of the species problem*, Madroño, vol. 12 (7), pp. 193-209.
- CAIN, A. J., 1954, *Animal species and their evolution* (Harper and Row, New York)
- DARWIN, C., 1859, *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (John Murray, London)
- DOBZHANSKY, T., 1951, *Genetics and the origin of species*, 3rd, ed. (Columbia University Press, New York)
- DOBZHANSKY, T., F. J. AYALA, G. LEDYARD STEBBINS and J. W. VALENTINE, 1977, *Evolution* (W. H. Freeman & Co., San Francisco)
- ELDREDGE, N., 1971, *The allopatric model and phylogeny in Paleozoic invertebrates*, Evolution, vol. 25, pp. 156-167
- ELDREDGE, N., and S. J. GOULD, 1972, *Punctuated equilibria: an alternative to phyletic gradualism*, in: *Models in Paleobiology*, ed. T. J. M. Schopf, pp. 82-115 (Freeman, Cope, and Co., San Francisco)
- ELDREDGE, N., and S. J. GOULD, 1978, *Evolutionary models and biostratigraphic strategies*, in: *Concepts and methods of biostratigraphy*, eds. E. G. Kaufman and J. E. Hazel, pp. 25-40 (Dowden, Hutchinson & Ross, Inc., Stroudsburg)
- FERRATER-MORA, J., 1962, *El ser y la muerte, bosquejo de filosofía integracionista* (Cited in FERRATER-MORA, 1966, Diccionario de Filosofía, I, pp. 938 (Editorial Sudamericana, Buenos Aires)
- FERRATER-MORA, J., 1966, *Realismo*, in: J. FERRATER-MORA, Diccionario de Filosofía, vol. II, pp. 538-540 (Editorial Sudamericana, Buenos Aires)
- GHISELIN, N. T., 1975, *A radical solution to the species problem*, Syst. Zool., vol. 23, pp. 536-544
- GILMOUR, J. S. L., 1961, *Taxonomy*, in: *Contemporary botanical thought*, eds. A. M. Mac-Leod and L. S. Cobley, pp. 27-47 (Oliver Boyd, Edinburg)
- GRANT, V., 1963, *The origin of adaptations* (Columbia University Press, New York, London)
- HANSON, E. D., 1961, *Animal diversity* (Prentice Hall Inc., Englewood Cliffs, N. Jersey)
- HULL, D. L., 1965, *The effect of essentialism on taxonomy: two thousand years of stasis*, Brit. J. Philos. Sci., vol. 15, pp. 314-326; vol. 16, pp. 1-18
- HULL, D. L., 1967, *The metaphysics of evolution*, Brit. J. Hist. Sci., vol. 3, pp. 309-337
- HULL, D. L., 1974, *Philosophy of biological sciences* (Prentice Hall Inc., Englewood Cliffs, N. Jersey)
- HULL, D. L., 1976, *Are species really individuals?*, Syst. Zool., vol. 25, pp. 174-191
- HULL, D. L., 1977, *The ontological status of species as evolutionary units*, in: *Foundational problems in the special sciences*, Part 2 of Proc. 5th Intern. Congress Log., Methodol. and Phil. of Sci., eds. R. E. Butts and J. Hintikka, pp. 91-102 (D. Reidel Publ., Dordrecht, Holland)

- IANKOVSKY, A. V., 1966, *The problem of integrity of biological species* (in Russian), in: Filosoficheskii problemy sovremennoi biologii, ed. S. Mamsin, pp. 155–176 (Izdatelstvo Nauka, Moscow)
- LAMARCK, J. B. de, 1907 (1806), *Discours d'ouverture An VIII, An X, An XI et 1806*, Bull. Scient. France et Belgique, p. 21
- LEHMAN, H., 1967, *Are biological species real?*, Phil. Sci., vol. 34, pp. 157–161
- IMBRIE, J., 1957, *The species problem with fossil animals*, in: The species problem, ed. E. Mayr, pp. 125–153 (Amer. Assoc. Adv. Sci. Publ. 50, Washington)
- MAYR, E., 1957, *Species concepts and definitions*, in: The species problem, ed. E. Mayr, vol. 1, pp. 22 (Amer. Assoc. Adv. Sci. Publ. 50, Washington)
- MAYR, E., 1963, *Animal species and evolution* (Harvard Univ. Press, Cambridge, Mass.)
- MAYR, E., 1969a, *The biological meaning of species*, Biol. J. Linn Soc., vol. 1, pp. 311–320
- MAYR, E., 1969b, *Principles of systematic zoology* (McGraw-Hill Book Co., New York)
- MEYER-ABICH A., 1948, *Naturphilosophie auf neuen Wegen* (Hippocrates Verlag, Stuttgart)
- MORGAN, G. I., 1933, *The emergence of novelty* (Williams & Norgate Ltd., London)
- PERRIER, E., 1884, *La philosophie zoologique avant Darwin* (Félix Alcan, éditeur, Paris)
- REIG, O. A., 1968, *Los conceptos de especie en la biología* (Caracas: Ediciones de la Biblioteca, Universidad Central de Venezuela)
- REIG, O. A., 1977, *Los diversos procesos de la emergencia de la novedad en la biología: su consistencia con la explicación y el método científico*, Simp. Interamer. Probl. Filosof. Biol. Contemp., pp. 1–21 (Soc. Venez. Philos., Caracas) (Mimeo.)
- ROBINET, J. B., 1768, *Considerations philosophiques sur la gradation naturelle des formes de l'être* (Paris)
- RUSE, M., 1969, *Definitions of species in biology*, Brit. J. Phil. Sci., vol. 20, pp. 97–119
- SETROV, M. I., 1966, *The meaning of Bertalanffy's general systems theory for biology* (in Russian), in: Philosophicheskii problemy sovremennoi biologii, ed. A. S. Mamsin, pp. 48–62 (Izdatelstvo Nauka, Moscow)
- SIMPSON, G. G., 1961, *Principles of animal taxonomy* (Columbia University Press, New York)
- SMART, J. J. C., 1963, *Philosophy and scientific realism* (Routledge & Kegan Paul, London)
- SMITH, H. M., 1958, *The synthetic natural populational species in biology*, Syst. Zool., vol. 7, pp. 116–119
- SONNEBORN, T. M., 1957, *Breeding systems, reproductive methods, and species problems in Protozoa*, in: The species problem, ed. E. Mayr, pp. 39–80 (Amer. Assoc. Adv. Sci. Publ. 50, Washington)
- SOUTHWOOD, T. R. E., 1976, *Bionomic strategies and population parameters*, in: Theoretical ecology, principles and applications, ed. R. M. May, pp. 26–48 (W. B. Saunders Co., Philadelphia & Toronto)
- VAN VALEN, L., 1976, *Individualistic classes*, Phil. Sci., vol. 43, pp. 539–541
- VAVILOV, N. I., 1931, *The Linnean species as a system* (in Russian), Trudy Prikl. Bot. Genet. Sci., vol. 26, pp. 109–134
- WATKINS, J. W. N., 1975, *Metaphysics and the advancement of science*, Brit. J. Phil. Sci., vol. 26, pp. 91–121.
- ZAVADSKY, K. M., 1966, *Principal organization forms of living organisms and their subdivisions* (in Russian), in: Philosophicheskii problemy sovremennoi biologii, ed. A. S. Mamsin, pp. 29–47 (Izdatelstvo Nauka, Moscow)
- ZIRKLE, C., 1959, *Species before Darwin*, Proc. Amer. Phil. Soc., vol. 103, pp. 639–644

SOME CONNECTIONS BETWEEN ASCIPTIONS OF GOALS AND ASSUMPTIONS OF ADAPTIVENESS*

ANDREW WOODFIELD

University of Bristol, Bristol, U.K.

In past ages and in remote cultures, all sorts of extraordinary thoughts and desires have been ascribed to non-human animals. In sixteenth century Europe, pigs were occasionally put on trial for heresy. Nowadays most people in Europe would regard it as anthropomorphic to ascribe Satanic desires to pigs. The prevailing view in the West seems to be that animals' actions can be 'tamed', so to speak, by being explained in terms of desires that arise out of basic biological needs. We let biological plausibility constrain our psychological speculation.

But it seems we do not apply the rule consistently. The bald principle 'Desires are ultimately grounded in needs' applies equally to all evolved creatures that have desires. Recent controversies in ethology and socio-biology, however, have revealed that where human behaviour is concerned there is sizeable opposition to the biological approach. But if it is unjustifiably *reductionist* to think that the goals of, say, an Amazonian Indian are all ultimately tailored to his or her biological needs, then might it not be equally unjustifiable to think the same thing about apes, dogs or pigs?

It is worth investigating in greater depth the modes of underpinning that biology is supposed to be able to provide for psychology. The present paper focusses on just one aspect of the problem. In what ways, and to what extent, can a *functional* understanding of animal behaviour provide a *grounding* for a theory of goals? This looks like a purely scientific question. But one needs to be very clear about what one means by 'grounding'. The kinds of grounding that are *possible* depend upon metascientific

* Thanks are due to David Hirschmann for his useful comments on the first draft of this paper.

considerations concerning the concept of *goal* and the form of functional explanations. To survey these possibilities in the abstract is really an exercise in the *philosophy* of psychobiology.

In the argument that follows, certain claims about functions and goals are going to serve as premisses. These assumptions cannot be fully argued for here, though I have tried to do so elsewhere (in *Teleology* (WOODFIELD, 1976), henceforth abbreviated to *Tel.*). The main assumptions had better be stated straightaway.

Concerning functionality, it is assumed that behaviour-patterns characteristic of members of a species are often *adaptive* in the sense that their performance helps the individual to survive and/or reproduce and/or be a good parent, without there being necessarily any recognition on the part of the behaving animal of these short or long-term benefits. Such behavioural tendencies are in many cases instinctive. They have evolved by natural selection in the same way that adaptive anatomical features have evolved. Indeed, every functional behaviour-disposition presupposes underlying physiological mechanisms, and that the animal should be so structured is itself an adaptation.

Where a characteristic behaviour-pattern is not instinctive, its presence in an individual could still perhaps be explained by a sort of ontogenetic analogue of natural selection: the behaviour 'survives' in so far as the animal is helped to survive through being a regular performer of it. Again, there need be nothing mentalistic involved in this process.

Whether the origin of the behavioural feature be phylogenetic or ontogenetic, to explain it *functionally* is to say that such behaviour tends to be exhibited under such and such conditions *because* that kind of behaviour confers an advantage in such conditions upon the individual, or the offspring, or the group, depending on the case, these being conditions that prevail (have so far prevailed) in the habitat in which the animals in question live. This is a characteristically biological form of explanation. The availability of a functional explanation marks off genuine adaptations from features that happen to confer a benefit fortuitously, on this or that occasion.

If the environment changes, a feature which was adaptive before the change may cease to be useful. Thus it is possible, during the aftermath of the change, for a feature to confer no present advantage even though its presence is explicable functionally. If so, we would probably continue to call it an adaptation out of deference to its origin. The distinctively biological approach to the understanding of behaviour *consists* largely

in the grasping of detailed facts concerning the organism's interactions with its environment which display the adaptedness of the former to the latter.

On the view of functional explanation sketched here, the concept of an adaptation has no logical connection with the concept of a goal. A functional explanation of a piece of behaviour says nothing about whether that behaviour is goal-directed. It may be or it may not be. It differs also from the view put forward by Professor NAGEL (1977), in which the notion of goal is used in *explicating* the notion of biological functional explanation. (Arguments against this view are proposed in *Tel.* pp. 124–130.)

Concerning goals, my primary assumption is that to ascribe a goal is to enter the realm of common-sense intentional psychology, with all its attendant explanatory apparatus. This is so because the term 'goal' means, in its core-sense, 'object of desire', and desiring is a psychological state.

A key feature of intentional psychology is that if you employ one of its proprietary terms, such as 'desire', you incur a commitment to employ various others, such as 'perceive', 'believe', and possibly 'learn', 'remember', and 'calculate'. These terms denote internal states or processes in the animal which are linked to one another in a complicated network, and are presumably 'realized' in computational states of the brain. The system of explaining behaviour psychologically as the net result of interactions between internal states is particularly efficient when the number of attributed states has to be very large. It is suited to explaining the behaviour of sophisticated animals that can learn new things and can cope with a complex changing environment.¹

Another key feature is that the states in question have intentional contents, specified by a 'that' clause or by an infinitive after the main verb in the sentence ascribing the state. Ascriptions of intentional attitudes credit the animal-subject with a capacity to model internally the

¹ Common usage allows psychological terms to be freely applied to gorillas, dogs, cats, rats, etc. Since I do not hold that the states denoted need necessarily be conscious, I have no qualms about applying the terms 'believe' and 'want' to many inframammalian animals. Those who are very strict in their deployment of mentalistic vocabulary may substitute other terms of their choosing, provided they are terms of a theory roughly isomorphic in *global architecture* (STICH, 1981) with folk psychology, and which take intentional content-clauses or infinitive verb complements. The term 'strive' could replace 'desire'. Jonathan Bennett's term 'register' is an excellent substitute for 'believe' (BENNETT, 1976, especially Sections 14, 15)

outside world as it appears from the animal's point of view. Such internal representations can be inaccurate. This fact correlates with the semantic fact that '*S* believes that *p*' is not a truth-function of '*p*'. Moreover, '*S* wants to do *G*' may be true even if it is never true that *S* does *G*. *S*'s desire may be thwarted. *G* may even be impossible for *S*.

To provide a formally satisfactory goal-explanation, one cites a desire and a belief whose content-specifications match up in the right way with the description of the action being explained. Both are needed. In the absence of one, the other does not explain why that act occurred. The form is: (Desire to do *G*) plus (Belief that *B* is a means to *G*) give rise to (behaviour *B*). In this triadic structure the term '*G*' appears twice, each time within an intentional content clause. The term "*B*" appears once within a content-clause, and once not within. This dual role for '*B*' guarantees a link between the way psychological states are classified and the way behaviour is classified. If two ethologists differ in the way they segment *S*'s acts, they will be forced to ascribe belief-states with different contents if they wish to explain those acts psychologically. Also, if one act can have two or more different descriptions, it can be goal-directed under one description, not goal-directed under another description. All these features are part of what is alluded to by the vague term 'intentionality'².

Let us briefly note some empirical facts about how the desires of an individual can be sorted and ordered. *S* may have a large number of desires at the same time. Also *S*'s desires may change over time. It is useful to introduce the notion of *S*'s *desire-profile at t* (or during *dt*), defined as the total set of desires possessed by *S* at a given time or over a given period.

Desires may be either instinctive or acquired. If *S* has desires at birth, these will count as instinctive. But a desire which makes itself felt for the first time in later life can also be instinctive. For instance, the urge to gather nestbuilding material and the urge to build a nest arise naturally at a certain stage in the life of a ring-dove as a result of hormonal changes induced by certain kinds of environmental stimuli. Although the dove needs to experience the stimuli, it is genetically programmed to respond physiologically in specific ways upon receipt of those stimuli at a certain stage of maturity.

² I set aside here the problem of the deviant causal pathway (see *Tel.* pp. 172-182; PEACOCKE, 1979), which raises doubts about whether desire-belief causation is *sufficient* for the intentional performing of *B*.

Acquired desires can arise in at least three different ways. First, and least importantly, a desire may conceivably be sparked off by some cause which has nothing to do with S 's normal functioning or maturation, such as a neurologist tampering with S 's neural circuitry. In such a case, there would be no reason to expect that the cause should have any connection with the object of desire, or that the desire should be useful to S .

Secondly, S may be rewarded with a pleasurable feeling or sensation while performing an act G idly without prior desire, and may then start to associate that act-type with pleasure. S becomes motivated to perform that act again under suitable circumstances. This phenomenon is standardly called 'The Law of Effect' (see BROADBENT, 1961, Chapters 2 and 7; DENNETT, 1975). The work of OLDS and others (1971) suggests that the experience of pleasure is supervenient upon neural activity in a number of sites in the medial forebrain bundle. When certain sites are electrically stimulated, the animal finds what it is doing at the time rewarding, and will tend to 'come back for more'. In this way an animal can be conditioned to want to do things which it previously did not want. In cases where S finds it naturally rewarding to do G , right from the first time it ever does G , it could be that S is genetically programmed to acquire the desire to do G .

The third way in which a new desire can arise is when S calculates that a certain act G_1 is a means to an act G_2 that it already wants to perform, and this calculation generates a subsidiary desire to do G_1 . Once a derived desire has become established in this way, it may become stable and entrenched. S may continue to want to do G_1 even when G_1 is no longer a means to G_2 . This mode of generating derived desires requires that S has some desire to start with. The initial desire could be acquired in either of the first two ways, or it could be instinctive.

Suppose that S has two desires at t . If they are independent of one another, they will vie for control over S 's current behaviour. But they need not be in competition. If S is sophisticated enough to acquire desires in the third way, he must be able to string actions into sequences under the guidance of a plan. Thus it may be that S wants to do G_1 , because S wants to do G_2 and believes that G_1 is a means to G_2 . This is an intra-psychological explanation of why S wants to do G_1 . A new question may then be raised: Why does S want to do G_2 ? If there is no explanation in terms of any further desire motivating this desire at t , let us call G_2 an *ultimate* goal of S 's at t . Of course, the new question can still be asked, and perhaps answered. But an explanation of why S has the ultimate

goal it has at t will be on a different level from before. It is at this point that the question of grounding the goal-theory in some other theory is most likely to arise. However, it would be wrong to think that such chains always terminate in instinctive desires, even if the *typical* role of instincts is to be cited as termini. (See DE SOUSA, 1979.) No restrictions have been placed on the content or origin of a desire that is ultimate for S at t . The notion of ultimacy at t is defined solely in terms of desire's causal role at t . A cryptogenic desire to eat mud is, in theory, just as capable of being ultimate as, say, a desire to mate.

Ways of grounding

Now that the terms have been defined, we are in a position to sort out all the ways in which it is possible to give a functional explanation of why S has such and such a desire or desire-profile in environment E at t . We know that such explanations can advert either to the phylogeny or the ontogeny of the desire. In either case, a necessary condition of the explanation's working is that the desire be of a type whose tokens reliably give rise (have given rise) to something biologically advantageous.

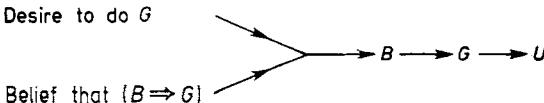
The first condition can be satisfied in many cases. If a certain behaviour-pattern or physiological process has biological utility, then any internal state of S which increases the probability of that behaviour or process will itself have utility, provided it does not at the same time increase the probability of something harmful to a degree which results in overall net disutility. A desire is just the sort of state that can be relied on, when coupled with an instrumental belief, to make behaviour more probable than it would otherwise have been. If squirrels need to eat nuts, and nuts are scarce in E , a squirrel which actively seeks nuts is more likely to come across some than a squirrel that merely waits for nuts to appear. The desire to find nuts helps the first squirrel to survive by motivating it to seek.

To complete the explanation, a further condition has to be met, that the desire be present in S as a characteristic *because* it satisfies the first condition. This presents no problems in principle. It is an empirical matter, although rather difficult to prove.

I now come to what is perhaps the main point of the paper. If I am right about the causal connections, it is evident that a desire to do G can have biological value for many reasons other than for the reason that it increases the likelihood that S will do G (where doing G is something that has biological value). The case where the desire's causal contribution

to some utility *goes via goal-achievement* is just one special case. In the special case, the chain of events is as follows. The desire to do *G* plus the belief that *B* is a means to *G* jointly give rise to *B*, which in turn contributes to the occurrence of *G*. *G* then goes on to contribute to some effect having biological utility, call it *U*.

Case (1) Desire to do G

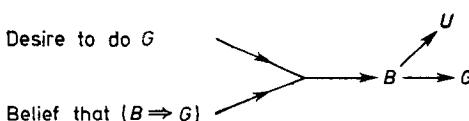


(Single arrows denote 'causal contributing' relation. Double arrow signifies 'is a means to'.)

It is worth noting that even in this case the utility of the concrete action described as '*G*' might not depend on its satisfying that description, that is, the description under which it counts as *S*'s goal. Suppose that the act *G* also satisfies another description, say '*G**'. If *S* had wanted to do *G**, *S* might have fulfilled this desire by performing the very same act, and the beneficial effect might have been the same. The usefulness of that act does not necessarily depend on its being of type *G*, the specific goal that it actually was. Indeed, its usefulness does not *depend on* its being any goal at all. If that same act had occurred for some reason without *S* wanting to perform it (under any of its descriptions), it might well have gone on to have the same beneficial effect. But it would not have counted as a goal of *S*'s.

That, then, is the first case. But there are indefinitely many other cases in which the causal chain bifurcates at a point prior to the achievement of *G*, such that the segment of the chain that leads to *U* does not go through *G*. One such case is where the goal-directed behaviour *B*, motivated by the desire and belief, makes its own contribution to *U*.

Case (2) Desire to do G

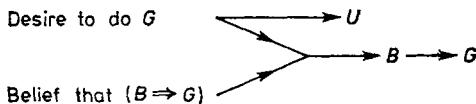


To take an example from human psychology: the desire to go to heaven plus the belief that you are more likely to get to heaven if you do not drink alcohol or smoke tobacco jointly motivate you not to drink or smoke. Your health benefits, whether or not you get to heaven. So the

desire to go to heaven has biological utility, as long as you believe the right things.

Another case is where the desire gives rise to some advantageous effect independently of its behaviourally motivating effect.

Case (3) Desire to do G



A possible example would be where the desire to fight off a potential intruder stimulates the release of adrenalin, which in turn tones up the system.

There are thus various possibilities, depending on where the bifurcation is located. Since in all cases other than (1) the utility of the desire does not stem from its producing *G*, the desire would still have utility even if *G* did not occur. The events on the *G*-line posterior to the bifurcation could be purely hypothetical. If the situation is as pure and simple as in cases (2) or (3), not only is it irrelevant to the functional explanation whether the goal is achieved, it is also irrelevant whether the desire increases the probability of the goal's being achieved. Not all desires do. *S*'s desire to jump over the moon does not make it more likely that *S* will jump over the moon. It is even irrelevant, in a way, what the goal *is*. Its only relevance is that it fixes the *identity* of the desire.

Most actual examples of functional animal desires will not be as pure as any of these cases. There will often be a mixture of beneficial effects resulting from different stages of the main process. Any reliably beneficial spin-off from an internal state could be responsible for its presence or persistence in *S*. When lion cubs play, for instance, the main functions of the activity appear to be to practise hunting-movements and to gain information about the environment. These functions would be equally well served by 'serious' attempts at hunting and exploring. The distinguishing mark of play is not its function, but the fact that it is done for fun. (See LORZOS, 1966, and SCHENKEL, 1966.) The cub who wants to pounce on the tail which his mother is flicking so invitingly may have acquired the desire simply through having gained pleasure from such activity in the past. Lion cubs seem to be innately disposed to find such pouncings pleasurable. That lions should be so programmed has functional utility in so far as it encourages the development of motor-patterns suited to adult leonine life. But it seems also true that fun in moderation itself

has utility in a sufficiently safe environment, because it reduces anxiety and promotes health.

An adaptive desire will normally derive its utility, or part of its utility, from being a member of an integrated set of desires. For instance, a rat needs to ingest so many grams per day of carbohydrate, protein and fat. Suppose that there is no single foodstuff around which contains the right balance, but there are three different kinds of food. If the rat eats the right proportion of each, it ends up well-nourished. It is advantageous to the rat to have separate desires for each of the three foods which wax and wane during the day in response to what has already been eaten. Each desire can be said to be adaptive, but only against the background of the other two (and the background of relevant beliefs).

Thus the relations between goals and functions are quite complicated. It is not true that all *S*'s desires are necessarily adaptive. It is not true that all non-adaptive desires must have arisen out of desires that were adaptive. It is theoretically possible that a bizarre individual's entire desire-profile at *t* be maladaptive, that *S* survives despite his will.

Desires that are adaptive derive their utility by various routes, some more indirect than others. It is not true that if a desire to do *G* is adaptive, doing *G* must also be adaptive. Nature, in its wisdom, has discovered that in many cases the optimal design for a system that will reliably achieve *U*, that is, will survive and reproduce, is to programme it so that it sometimes tries not for *U* but for *G*.

Can biology help us to identify *S*'s goals?

The discussion so far has assumed that the desires (and beliefs) of the organism are fixed and known. The task has been to tease out the various ways in which a given desire or set of desires *can be* adaptive. I have said nothing about how to construct a desire-profile for *S*, nor about how to justify one's hypotheses about *S*'s psychological states. Let us now look at the situation from the point of view of an observer who is trying to work out what an animal or human wants. Can any principles concerning adaptiveness guide him in this task?

There is little doubt that ethologically minded psychologists do, in fact, employ biological assumptions as touchstones at various points when they put forward psychological hypotheses. And this practice has filtered through a bit into everyday thinking about animals. But it could be that contemporary practice reflects a passing fashion, a biologicistic

ideology which has little rational justification. Or, it could be that our thinking is still biologically unsophisticated and that we do not do it nearly enough. So let us ask, in the light of the previous findings, to what extent ought psychology to be *methodologically* linked to biological functionalism?

Theory construction in intentional psychology has, of course, rules of its own that are autonomous of biology or any other theory. Since the point of psychological hypotheses is to predict and explain behaviour, it is to behaviour that the theory looks for its data. However, it is widely remarked that the totality of behavioural data available up till t underdetermines the subject's psychological profile at t . Any theory that makes falsifiable predictions is underdetermined by the data on which it rests at t . But the situation is worse in psychology, because the subject changes over time. The theory has to take account of changes in S 's psychological profile, some of which are caused by events that reflect an earlier profile. Not only are there various alternative desire-belief profiles for S at t which explain S 's actual behaviour equally well, it is also the case that some of these profiles can never be ruled out as false by the evidence of what S does later. They can be kept in play, provided they are supplemented by a suitable hypothesis about how S has changed.

How is one to choose, then, between different hypotheses about what S wants? Although it seems clear that some hypotheses are going to be more plausible than others, it is not clear what plausibility is. Many would argue that plausibility should be judged partly by reference to biological facts.

Professor D. C. Dennett, for example, argues that if we are treating S as an Intentional System, we should ascribe desires according to the following principle: 'A system's desires are those *it ought to have*, given its biological needs and the most practicable means of satisfying them' (DENNETT, 1979, p. 10).³ Professor Jonathan Bennett, on the other hand, would probably argue that behaviour is the only basis for a theory of goals. He thinks that the concept of biological need is not needed in goal-theory (BENNETT, 1976, p. 62). The consequence of the latter view

³ Dennett has another such principle for ascribing *beliefs*: 'A system's beliefs are those *it ought to have*, given its perceptual capacities, its epistemic needs and biography'. It should perhaps be said that I am evaluating the desire-principle merely as a useful rule of thumb. Dennett goes further than this. He holds that these principles are partially definitive of the notions of desire and belief as used in 'Intentional System Theory'.

is that if there were two rival psychological theories for *S*, both equally supported by all the behaviour so far, there would be no basis for choosing between them until further behavioural evidence came in that could decide the matter.

At first sight, it looks as though Dennett's view has the edge, at least as a heuristic policy. If there is a functional explanation of why *S* should have the desire-profile postulated by one theory, but no explanation of why *S* should have the desires postulated by a second theory, then even though it is perfectly conceivable that theory (2) is right, and that it might later be shown to be right, it is more rational to accept theory (1) for the time being as a working hypothesis. If a theory of needs can provide a grounding for theory (1), then in the absence of any other background theory which tells against the biological theory, psychology should gratefully accept the support that biology offers.

Suppose that the two theories make different predictions which can be tested experimentally. Then, of course, the result of the test should decide. But the heuristic principle said that the more adaptive desire-profile was *more likely* to be the true one. This principle may itself turn out to be inductively supported as evidence mounts up of cases where the theory which assigns the more adaptive desire-profile explains the subsequent behavioural evidence better than its rivals.

But let us take a closer, more quizzical, look. Everything turns on our having a good understanding of what desires *S* *ought to have*, given its needs. If I am right about the multiplicity of ways that desires *can be* adaptive, the injunction to select those which there is good reason to think *are* adaptive will be unhelpful and unusable in many cases. It does not take much imagination to invent a story showing how pretty well any desire you like to mention could prove useful to *S* through some circuitous route. To prevent the 'biological grounding' principle from degenerating into vacuousness, we must ensure that our story is not merely plausible, but probably true. This means that we need to be in possession of lots of factual information about the likely effects of that desire upon *S*'s life. If *S* is a member of a well-studied species whose behaviour is fairly stereotyped, largely instinctive, and where the members do not exhibit many individual differences, the assessment of utility might not be too difficult. But the calculations become terribly complicated when *S* is a higher mammal having a complex cognitive structure, because there are so many channels through which a desire can have effects. The biological touchstone becomes practically unusable, since we cannot work

out which desires would be the most adaptive ones for S to have (beyond those at the most rudimentary level).

Secondly, the theoretical justification for using the biological grounding rule becomes progressively weaker the higher S is on the evolutionary scale. The rationale for having a general preference for an 'adaptive desire' hypothesis over a 'non-adaptive desire' hypothesis is that, if you were in a position to pick at random any desire from S 's total set at t , the chances are that you would pick one that is adaptive. It is presumed, in other words, that most of S 's desires are sure to be adaptive. For a creature whose desire-profile contains a high proportion of acquired desires, there is much less justification for making such a presumption. None of the three ways of acquiring a desire guarantee that the new desire will be functionally good. The S we are dealing with might happen to be an animal whose acquired tastes are predominantly contra-biological. The policy of ascribing to S the desires S ought to have would lead a naïve psychologist into disastrous error.

Thirdly, the grounding rule is supposed to help us endow S 's desires with the right specific contents. But it is actually incapable of doing anything so definite as that, even when the desire whose content is in question is known to have useful consequences. As we have seen, the utility of a desire need not come from the utility of its *satisfaction*. The rule cannot be construed as saying: Pick the goal whose achievement serves S 's needs best. In all cases other than the special case (1), achieving the goal is, so to speak, the icing on the cake.

Assume, for the sake of argument, that the main biological function of sports such as rugby, tennis, soccer and cricket is exercise. The main goals of a certain sportsman S in playing tennis are to win and to have fun. His desire to win motivates him to play hard, consequently he gets a lot of exercise. His desire is adaptive even if he does not manage to win. The *notion* of winning is the pivot around which the entire rule-governed pattern of activity is structured; but *actually* winning is a useless bonus as far as biology is concerned.

The main biological function of mating, to take another example, is the production of offspring. Few animals are driven to mate by a desire for offspring, though. What they want is to have sex. In this case, achieving the goal is a necessary intermediate step on the way to the biological end. But for human beings at least, there are various ways of having sex. To each position listed in the *Kama Sutra* there corresponds a distinct possible human desire to have sex in that position. Each one has functional utility

in the same way as long as it leads to mating. (Set aside the fact that successful mating is highly unlikely in some of the positions.) But if your goal is to have sex while sitting in the Lotus position, then achieving *that* is an act of supererogation as far as biology is concerned.

It is clear, then, that there will always be a *range* of desire-hypotheses to explain *S*'s behaviour with respect to which the biological grounding rule is indifferent, because all the desires within that range have the same utility by the same causal route. Behaviour underdetermines psychology, true enough, but it is also true that behaviour *plus biological grounding* underdetermine psychology. On reflection, this is hardly to be wondered at. Type-identifying states by their contents is the finest-grained taxonomy we know. All other modes of grouping are gross by comparison. It is bound to be the case that some desires having distinct contents will be functionally equivalent.

My final worry is that over-zealous deployment of the 'biological grounding' idea might lead to the neglect of principles of testing proper to psychological theories. The enthusiastic psychobiologist runs the risk of falling into the following bad habit. He notices that on various occasions the animal performs sequences of apparently goal-directed movements which have a certain outcome *F*, and *F* has survival-value. He hypothesizes that *F* is the goal of the behaviour. He then tests this hypothesis under varying conditions, and finds that the animal's behaviour shows a high degree of persistence and plasticity with respect to *F*. This is taken as confirmation of the hypothesis. Satisfied with this success, he publishes his paper. His findings are hailed and cited in later textbooks. Unfortunately, however, he neglected to consider a number of alternative hypotheses, one being the hypothesis that the goal of the animal is *G*. Under all the conditions he considered, *F* and *G* were compresent. Crucial tests could have been performed under conditions in which the animal's doing *F* would preclude its doing *G*, and under conditions in which doing *G* would preclude doing *F*. If only these tests had been carried out, the scientist would have seen that the aim of the activity from the animal's point of view was not to do *F* at all. *F* was a useful by-product that happened always to occur, under the initial range of conditions. The animal was not responding in ways that were sensitive to information available to it about the likelihood that these responses would help to bring about *F*.

A theory of goals cannot explain behaviour *B* just by itself. It has to go hand in hand with a theory of the animal's beliefs. To test a goal-hypothesis properly, one has to test also whether the animal believes

that *B* is a means to that goal.⁴ Does it have access to the relevant information? To what aspects of the stimulus-situation is it sensitive? In what ways can it be fooled? To fail to perform such tests is to forget that the goal-description is locked inside an intentional content-clause, and that what *counts* as a goal must be relativized to *S*'s point of view.

Psychological theorizing is doomed to live with this constraint of intentionality. Nevertheless, I do not wish to *exaggerate* the limitations that it imposes upon our choice of words when we try to describe what an animal's goal is. Ascriptions of desires which have content-clauses containing definite singular terms can be construed *transparently* with respect to those terms. Thus, if I ascribe to my dog a desire to eat the steak on my plate, I need not be implying that the dog thinks of it as *steak*. But I do imply that he picks out the steak as an object, that he responds to sensory stimuli that in fact have the steak as their source, and that he thinks of it as being something nice to eat. (Cf. ARMSTRONG, 1973, pp. 25-27; see also STICH, 1979.)

Another way, therefore, in which a desire-hypothesis can be given the *appearance* of being biologically grounded is to describe the object of desire in technical biological vocabulary, but to construe the description transparently. Instead of saying 'The dog wants to eat the steak', one might say 'The dog wants to ingest the substance which contains 63% of its daily protein requirement'. This is harmless scientific pedantry, given that both utterances ascribe the same desire. One is free to select the description which is best suited to the communicatory purpose in hand.

It is often possible to convey information about the *functional* role of the object by choosing a suitable description. Thus if a pregnant rat is placed in a cage that contains some cotton wool, the observer may explain the rat's cotton-wool oriented behaviour by saying 'She wants to collect the nest-building material'. He hints at what she is going to

⁴ In Professor Bennett's words, 'It is as though we had to test the freedom from distortion of a collection of panes of glass, under the restriction that we must always look through two panes at once.' (BENNETT, 1976, p. 51.) Bennett provides a detailed account of the indirect method of testing goal-hypotheses. Nearly all of what he says about testing-methods in sections 15 and 16 can be accepted even if you think that his *analysis* of 'goal' is mistaken. Having introduced the epistemic concept of registration, Bennett does not think it necessary to postulate a separate concept for a correlative motivational state. For him, goals are not explicated as being desired prospects against which the contents of registrations can be *matched* by *S*, but are defined through the notion of a teleological law relating registrations to actions. These matters are rather complicated, and really require a separate paper.

do with the cotton-wool once she has got it. But this has more to do with the pragmatics of content-ascription than with the *grounding* of psychology in biology.

In conclusion, I feel that little is to be gained by mixing up biological assumptions with psychological theorizing. The richer the psychology of the animal, the less there is to be gained. It merely obscures the central truth that the ascribing of contentful psychological states is an essentially egomorphic enterprise. The task is to capture in my language what the world looks like from the point of view of another subject. The best methodology I can use is to employ my intuition to the full and try to project myself into *S*'s position, having found out as much information as possible about *S*'s lifestyle, habits and capacities. Ethologically trained observers are better at this than people who know nothing of the biology of the animal, no doubt. But the effect of the training is to sharpen their intuition. It has this effect not just by improving their theoretical understanding (though this surely helps), but by turning them into people who *understand* animals better.

Perhaps there is even a sort of biological justification for using sympathetic intuition in psychology. In *Animal Life and Intelligence*, C. LLOYD MORGAN (1890) claimed that psychological understanding was not confined to humans. On page 340 he wrote, 'For myself, I cannot doubt that animals project into each other the shadows of the feelings of which they are themselves conscious'. Whether or not he was right about this, it is certainly true that humans can do it, and that this ability is a useful feature that could have arisen through natural selection. Lower animals are, in differing degrees, like human beings; we share ancestors. Many kinds of animals have evolved in close proximity to human beings; each species has been part of the ecology of the others. It seems likely, therefore, that humans are naturally programmed to guess right a lot of the time about what other animals want. I conjecture that Martians would not do so well, even if they had all the same prior facts to hand. One thing seems clear anyway: there is no *algorithm* for selecting the best goal-hypothesis.

References

- ARMSTRONG, D. M., 1973, *Belief, truth and knowledge* (Cambridge)
BENNETT, Jonathan, 1976, *Linguistic behaviour* (Cambridge)
DENNETT, D. C., 1975, *Why the law of effect will not go away*, Journal of the Theory of Social Behaviour, vol. 2. Reprinted as chapter 5 of *Brainstorms* (Bradford Books and Harvester, Sussex 1978)

- DENNETT, D. C., 1979, *Three kinds of intentional psychology*, forthcoming in a Thyssen Philosophy Group volume to be edited by R. A. Healey.
- BROADBENT, D. E., 1961, *Behaviour* (London)
- LOIZOS, Caroline, 1966, *Play in mammals*, in: *Play, Exploration and Territory in Mammals, Symposia of the Zoological Society of London*, No. 18, eds. P. A. Jewell and Caroline Loizos (London)
- MORGAN, C. Lloyd, 1890, *Animal life and intelligence* (Edward Arnold, London 1890-1)
- NAGEL, E., 1977, *Teleology revisited*, Journal of Philosophy, vol. LXXIV
- OLDS, J., W. S. ALLAN, and R. BRIESE, 1971, *Differentiation of hypothalamic drive and reward centers*, American Journal of Physiology, vol. 221, pp. 368-375
- PEACOCKE, Christopher, 1979, *Holistic explanation* (Oxford)
- SCHENKEL, Rudolf, 1966, *Play, exploration and territoriality in the wild lion*, in: Jewell and Loizos (eds). See under Loizos
- SOUSA, R., de, 1979, *Instinct and teleology* (as yet unpublished)
- STICH, Stephen P., 1979, *Do animals have beliefs?*, Australasian J. Phil., vol. 57
- STICH, Stephen P., 1981, *On the ascription of content*, in: *Thought and object*, ed. Andrew Woodfield (Oxford)
- WOODFIELD, Andrew, 1976, *Teology* (Cambridge)

FUNCTIONALISM

NED J. BLOCK

MIT Department of Linguistics and Philosophy, Cambridge, MA, U.S.A.

Functionalism

It is doubtful whether doctrines known as “functionalism” in fields as disparate as anthropology, literary criticism, psychology, and philosophy of psychology have anything in common but the name. Even in philosophy of psychology, the term is used in a number of distinct senses. The functionalisms of philosophy of psychology are, however, a closely knit group; indeed, they appear to have a common origin in the works of Aristotle (see HARTMAN, 1977, especially Ch. 4).

Three functionalisms have been enormously influential in philosophy of mind and psychology:

(1) *Functional analysis*. In this sense of the term, functionalism is a type of explanation, and, derivatively, a research strategy, the research strategy of looking for explanations of that type. A functional explanation is one that relies on a decomposition of a system into its component parts; it explains the working of the system in terms of the capacities of the parts and the way the parts are integrated with one another. For example, we can explain how a factory can produce refrigerators by appealing to the capacities of the various assembly lines, their workers and machines, and the organization of these components. (See CUMMINS, 1975, FODOR, 1965, 1968a, 1968b, and DENNETT, 1975.)

(2) *Computation-representation functionalism*. In this sense of the term, “functionalism” applies to an important special case of functional explanation as defined above, namely, to psychological explanation seen as akin to providing a computer program for the mind. Whatever mystery

our mental life may initially seem to have been dissolved by functional analysis of mental processes to the point where they are seen to be composed of computations as mechanical as the primitive operations of a digital computer—processes so stupid that appealing to them in psychological explanations involves no hint of question-begging. The key notions of functionalism in this sense are representation and computation. Psychological states are seen as systematically representing the world via a language of thought, and psychological processes are seen as computations involving these representations. See FODOR (1975).

(3) *Metaphysical functionalism.* The last functionalism, the one that this paper is mainly about, is a theory of *the nature of the mind*, rather than a theory of psychological explanation. Metaphysical functionalists are concerned not with how mental states account for behavior, but rather with what they *are*. The functionalist answer to "What are mental states?" is simply that mental states are functional states. Thus, theses of metaphysical functionalism are sometimes described as functional state identity theses. The main concern of metaphysical functionalism is the same as that of behaviorism and physicalism. All three doctrines address themselves to such questions as "What is pain?"—or at least to "What is there in common to all pains in virtue of which they are pains?"

It is important to note that metaphysical functionalism is concerned (in the first instance) with mental state *types*, not tokens, with *pain* for instance and not with particular pains. Most functionalists are willing to allow that each *particular* pain is a physical state or event, and indeed that for each type of pain-feeling organism, there is (perhaps) a single type of physical state that realizes pain in that type of organism. Where functionalists differ with physicalists, however, is with respect to the question of what is in common to all pains in virtue of which they are pains. The functionalist says the something in common is functional, while the physicalist says it is physical (and the behaviorist says it is behavioral).¹ Thus, in one respect, the disagreement between functionalists

¹ Discussions of functional state identity theses have sometimes concentrated on one or another weaker theses in order to avoid issues about identity conditions on entities such as states or properties (e.g. BLOCK and FODOR, 1972).

Consider the following theses:

- (1) Pain = functional state *S*.
- (2) Something is a pain just in case it is a (token of) *S*.

and physicalists (and behaviorists) is *metaphysical without being ontological*. Functionalists can be physicalists in allowing that all the entities (things, states, events, etc.) that exist are physical entities, denying only that what binds certain types of things together is a physical property.

Metaphysical functionalists characterize mental states in terms of their causal roles; particularly, in terms of their causal relations to sensory stimulations, behavioral outputs, and other mental states. Thus, for example, a metaphysical functionalist theory of pain might characterize pain in part in terms of its tendency to be caused by tissue damage, by its tendency to cause the desire to be rid of it, and by its tendency to produce action designed to separate the damaged part of the body from what is thought to cause the damage.

What I have said about metaphysical functionalism so far is rather vague, but, as will become clear, disagreements among metaphysical functionalists preclude easy characterization of the doctrine. Before going on to describe metaphysical functionalism in more detail, I shall briefly sketch some of the connections among the functionalist doctrines just enumerated. One connection is that functionalism in all the senses described has something to do with the notion of a Turing machine. (If you do not know what a Turing machine is, read the first paragraph of the next section of this paper.) Metaphysical functionalism often identifies mental states with Turing machine "table states" (as described in the next section). Computation-representation functionalism sees psychological explanation as something like providing a computer program for the mind. Its aim is to give a functional analysis of mental capacities into their component mechanical processes. If these mechanical processes are *algorithmic*, as is sometimes assumed (without much justification, in my view) then they will be Turing-computable as well (as the Church-Turing thesis assures us).² Functional analysis, however, is concerned with the notion of

(3) The conditions under which x and y are both pains are the same as the conditions under which x and y are both tokens of S .

(1) is a full-blooded functional state identity thesis which entails (2) and (3). Theses of the form of (2) and (3) can be used to state what it is that all pains have in common in virtue of which they are pains.

² DENNETT (1975) and REY (1979) make this appeal to the Church-Turing thesis. But if the mechanical processes involved analog rather than digital computation, then the processes could fail to be algorithmic in the sense required by the Church-Turing thesis. The experiments reported in the imagery section of BLOCK (1979) suggest that mental images are (at least partially) analog representations, and that the computations that operate on images are (at least partially) analog operations.

a Turing machine mainly in that providing something like a computer program for the mind is a special case of functional analysis.

Another similarity among the functionalisms mentioned is their relation to physical characterizations. The causal structures with which metaphysical functionalism identifies mental states are realizable by a vast variety of physical systems. Similarly, the information processing mechanisms postulated by a particular computation-representation functionalist theory could be realized hydraulically, electrically, or even mechanically. Finally, functional analysis would normally characterize a manufacturing process abstractly enough to allow a wide variety of types of machines (wood or metal, steam-driven or electrical), workers (human or robot or animal), and physical setups (a given number of assembly lines or half as many dual-purpose assembly lines). A third similarity is that each type of functionalism described legitimates at least one notion of functional equivalence. For example, for functional analysis, one sense of functional equivalence would be: has capacities that contribute in similar ways to the capacities of a whole.

In what follows, I shall try to give the reader a clearer picture of metaphysical functionalism. ("Functionalism" will mean metaphysical functionalism in what follows.)

Machine versions of functionalism

Some versions of functionalism are couched in terms of the notion of a Turing machine, while others are not. A Turing machine is specified by two functions: one from inputs and states to outputs, and one from inputs and states to states. A Turing machine has a finite number of states, inputs, and outputs, and the two functions specify a set of conditionals, one for each combination of state and input. The conditionals are of this form: if the machine is in state S and receives input I , it will then emit output O and go into next state S' . This set of conditionals is often expressed in the form of a machine table (see below). Any system that has a set of inputs, outputs, and states related in the way specified by the machine table is *described* by the machine table, and is a *realization* of the abstract automaton specified by the machine table. (This definition actually characterizes a finite automaton, which is just one kind of Turing machine.)

One very simple version of machine functionalism (see PUTNAM, 1967, and BLOCK and FODOR, 1972) states that each system that has mental

states is described by at least one Turing machine table of a certain specifiable sort: it also states that each type of mental state of the system is identical to one of the machine table states specified in the machine table. Consider, for example, the Turing machine described in the following "coke machine" machine table (cf. NELSON, 1975):

| | S_1 | S_2 |
|-----------------|-------------------------------|---|
| nickel input | Emit no output Go to S_2 | Emit a coke Go to S_1 |
| dime input | Emit a coke Stay in S_1 | Emit a coke and a nickel Go to S_1 |

One can get a crude picture of the simple version of machine functionalism described above by considering the claim that S_1 = dime-desire, and S_2 = nickel-desire. Of course, no functionalist would claim that a coke machine desires anything. Rather, the simple version of machine functionalism described above makes an analogous claim with respect to a much more complex machine table.

Machine versions of functionalism are useful for many purposes, but they do not provide the most general characterization of functionalism. One can achieve more generality by characterizing functionalism as the view that what makes a pain a pain (and what makes any mental state the mental state it is) is its having a certain causal role.³ But this formulation buys generality at the price of vagueness. A more precise formulation can be introduced as follows.⁴ Let T be a psychological theory (of either common sense or scientific psychology) that tells us (among other things) the relations among pain, other mental states, sensory inputs, and behavioral outputs. Reformulate T so that it is a single conjunctive sentence with all mental state terms as singular terms—e.g., 'is angry' becomes 'has anger'. Let T so reformulated be written as

$$T(s_1 \dots s_n)$$

³ Strictly speaking, even the causal role formulation is insufficiently general, as can be seen by noting that Turing machine functionalism is not a special case of causal role functionalism. Strictly speaking, none of the states of a Turing machine need cause any of the other states. All that is required for a physical system to satisfy a machine table is that the counterfactuals specified by the table are true of it. This can be accomplished by some causal agent outside the machine. Of course, one can always choose to speak of a *different* system, one that includes the causal agent as part of the machine, but that is irrelevant to my point.

⁴ Formulations of roughly this sort were first advanced by LEWIS (1966, 1970, 1972), MARTIN (1966). See also BLOCK (1978), FIELD (1978), GRICE (1975), and HARMAN (1973).

where s_1, \dots, s_n are terms that designate mental states. Replace each mental state term with a variable and prefix existential quantifiers to form the Ramsey sentence of the theory

$$\exists x_1 \dots x_n T(x_1 \dots x_n).$$

Now if x_i is the variable that replaced ‘pain’, we can define ‘pain’ as follows:

$$y \text{ has pain if and only if } \exists x_1 \dots x_n [T(x_1 \dots x_n) \& y \text{ has } x_i].$$

That is, one has pain just in case he has a state that has certain relations to other states that have certain relations to one another (and to inputs and outputs; I have omitted reference to inputs and outputs for the sake of simplicity). It will be convenient to think of pain as the property expressed by the predicate ‘ x has pain’, that is, to think of pain as the property ascribed to someone in saying that he has pain.⁵ Then, relative to theory T , pain can be identified with the property expressed by the predicate

$$\exists x_1 \dots x_n [T(x_1 \dots x_n) \& y \text{ has } x_i].$$

For example, take T to be the ridiculously simple theory that pain is caused by pin pricks and causes worry and the emission of loud noises, and worry, in turn, causes brow wrinkling. Relative to this simple theory, pain can be identified with the property expressed by the following predicate:

$$\exists x_1 \exists x_2 [(x_1 \text{ is caused by pin pricks and causes } x_2 \text{ and emission of loud noises} \& x_2 \text{ causes brow wrinkling}) \& y \text{ has } x_1].$$

That is, pain is the property that one has when one has a state that is caused by pin pricks, and causes emission of loud noises, and also causes something else, that, in turn, causes brow wrinkling.

We can make this somewhat less cumbersome by letting an expression of the form ‘ $\lambda x Fx$ ’ be a singular term meaning the same as an expression of the form ‘the property of being an x such that x is F ’, that is, ‘being F ’. So $\lambda x(x \text{ is bigger than a mouse} \& x \text{ is smaller than an elephant}) =$ being bigger than a mouse and smaller than an elephant.

Using this notation, we can say

$$\begin{aligned} \text{pain} = \lambda y \exists x_1 \exists x_2 & [(x_1 \text{ is caused by pin pricks and causes } x_2 \\ & \text{and emission of loud noises} \& x_2 \text{ causes brow wrinkling}) \\ & \& y \text{ has } x_1], \end{aligned}$$

⁵ See FIELD (1978) for an alternative convention.

rather than saying that pain is the property expressed by the predicate

$$\lambda x_1 \lambda x_2 [(x_1 \text{ is caused by pin pricks and causes } x_2 \text{ and emission of loud noises} \& x_2 \text{ causes brow wrinkling}) \& y \text{ has } x_1].$$

It may be useful to consider a non-mental example. It is sometimes supposed that automotive terms like ‘valve-lifter’ or ‘carburetor’ are functional terms. Anything that lifts valves in an engine with a certain organizational structure is a valve-lifter. (‘Camshaft’, on the other hand, is a “structural” term, at least relative to ‘valve-lifter’; a camshaft is *one* kind of device for lifting valves.)

Consider the “theory” that says: “The carburetor mixes gasoline and air and sends the mixture to the ignition chamber, which in turn...”. Let us consider ‘gasoline’ and ‘air’ to be input terms, and let x_1 replace ‘carburetor’, and x_2 replace ‘ignition chamber’. Then the property of being a carburetor would be

$$\lambda y \lambda x_1 \dots x_n [(\text{The } x_1 \text{ mixes gasoline and air and sends the mixture to the } x_2, \text{ which, in turn...}) \& y \text{ is an } x_1].$$

That is, being a carburetor = being what mixes gasoline and air and sends the mixture to something else, which, in turn...

This identification, and the identification of pain with the property one has when one is in a state that is caused by pin pricks and causes loud noises and also causes something else that causes brow wrinkling, would look less silly if the theories of pain (and carburetion) were more complex. But the essential idea of functionalism, as well as its major weakness, can be seen clearly in the example, albeit rather starkly. Pain is identified with an abstract causal property tied to the real world only via its relations, direct and indirect, to inputs and outputs. The weakness is that it seems so clearly conceivable that something could have that causal property, yet *not be* a pain. This point can be made more vivid by attention to the following example.

Imagine a body externally like a human body, say yours, but internally quite different. The neurons from sense organs are connected to a bank of lights in a hollow cavity in the head. The motor output neurons are activated by buttons on a console in another section of the cavity. On one wall of the cavity is a very large machine table that describes a person, say you, and in a corner of the cavity, there is a blackboard. In the cavity resides a little man. We tell the man to “start” in a certain state, say $S_{1,975}$,

which happens to be the state that you are now in. The man's job is as follows. He writes '1,975' on the blackboard, and looks to the bank of input lights to see what inputs are currently occurring. He sees a pattern of lights that he has been trained to identify as input I_{342} . He looks on the great machine table on the wall for the box at the intersection of row 342 and column 1,975. In the box is written "go to $S_{7,651}$; emit output $O_{10,283}$ ". The little man erases the '1,975' from the blackboard and writes '7,651'. Then he goes to the output console and presses the pattern of buttons that he has been taught will produce output $O_{10,283}$. Then he looks to the input board to see what the next input is, and so on. Through the efforts of this little man, the artificial body in which he lives behaves just as you would, given any possible sequence of inputs (starting now, since he has now been "set" to your current functional state—were the two of you to receive different inputs, your behavior might diverge). Indeed, since the machine table that the little man has posted in his cavity describes you, the homunculus-headed system is functionally identical to you.

But though the homunculus-headed system is now in the same functional state that you are in, does it have the same mental states that you have? Suppose that you and your homunculus-headed doppelganger are in different rooms, both connected to a third room by a two-way TV system. An interrogator in the third room addresses a remark carried to you by your TV set and also to your homunculus-headed doppelganger by his TV set. Since you and your doppelganger are functionally alike, you emit exactly the same sounds and movements in response. The interrogator replies, and you and your doppelganger continue to respond in indistinguishable manners. Of course, *you* are understanding the interrogator's English sentences and expressing your thoughts in English. But is your doppelganger understanding the interrogator's English sentences? Is it having thoughts? The *little man* need not understand the interrogator's remarks. He can push the buttons and read the patterns of lights without having any idea that they have any relation to a conversation in English. Indeed, he can do his job without knowing what sort of a system he is controlling, and without being able to understand English at all. Since the little man need not understand English, why should we suppose that the system consisting of him plus the body he controls must understand English?

Further, if you have a nasty headache and are calling for aspirin, and the homunculus-headed system is in the same functional state as you

(and is uttering similar sounds), surely there is reason to doubt whether it has pain (or indeed whether it has any qualitative state at all).

Now while such examples may be compelling, they should not convince anyone that functionalism is false. Intuitions about such matters are easy to manipulate, and even if our intuitions in these cases were immutable, a prudent philosopher should demand that they be augmented by argument. See BLOCK (1978) for an attempt to provide some arguments to this effect, and see LYCAN (forthcoming) for a critique. SHOEMAKER (1975) attempts to defend functionalism from a variety of "inverted qualia" and "absent qualia" objections, and BLOCK (forthcoming) replies.

Functionalism and behaviorism

Many functionalists (e.g., Lewis, Armstrong, Smart) consider themselves descendants of behaviorists who attempted to define a mental state in terms of what behaviors would be emitted in the presence of specified stimuli. For example, the desire for an ice-cream cone might be identified with a set of dispositions, including the disposition to reach out and grasp an ice-cream cone if one is proffered, other things being equal. But, as functionalist critics have emphasized (see PUTNAM, 1963; the point dates back at least to CHISHOLM, 1957, Chapter II, and GEACH, 1957, p. 8), the phrase "other things being equal" is behavioristically illicit, because it can only be filled in with references to *other mental states*. One who desires an ice-cream cone will be disposed to reach for one only if he *knows* it is an ice-cream cone (and not, in general, if he believes it to be a tube of axle-grease), and only if he does not *think* that taking an ice-cream cone would conflict with *other desires* of more importance to him (e.g., the desire to lose weight, avoid obligations, or avoid cholesterol). The final nail, in the behaviorist coffin was provided by the well-known "perfect actor" family of counterexamples. As PUTNAM (1963) argued in convincing detail, it is possible to imagine a community of perfect actors who, in virtue of lawlike regularities, have exactly the behavioral dispositions envisioned by the behaviorists to be associated with absence of pain, even though they do in fact have pain. This shows that no behavioral disposition is a necessary condition of pain, and an exactly analogous example of perfect pain-pretenders shows that no behavioral disposition is a sufficient condition of pain, either.

Functionalism in all its forms differs from behaviorism in two major respects. First, while behaviorists defined mental states in terms of stimuli

and responses, they did not think mental states were *themselves* causes

Shoemaker) would qualify as versions of behaviorism (since all of the original mental state terms are replaced by variables in the Ramsey sentence). Many other definitions of "behaviorism" count functionalism as a type of behaviorism. But it would be ludicrously literal-minded to take such definitions very seriously. Clear and general formulations of functionalism were not available until recently, so standard definitions of behaviorism could hardly be expected to draw the boundaries between behaviorism and functionalism with perfect accuracy. Furthermore, given an explicit definition of behaviorism, logical ingenuity can often disguise a functionalist account so as to fit the definition. (See THOMAS, 1978; and BEALER, 1978, for accomplishments of this rather dubious variety.) Definitions of behaviorism that count functionalism as behaviorist are misguided precisely *because* they blur the distinctions between functionalism and behaviorism just sketched. A characterization of pain can hardly be counted as behaviorist if it allows that a system could behave (and be disposed to behave) exactly *as if* it were in pain in all possible circumstances, yet *not be* in pain.⁶

Is functionalism reductionist?

Functionalists sometimes formulate their claim by saying that mental states can only be characterized in terms of other mental states. For instance, a person desires such and such if he would do so and so if he believed doing so and so will get him such and such, and if he believed doing so and so would not conflict with other desires. This much functionalism brings in no reductionism. But functionalists have rarely stopped there. Most regard mental terms as eliminable *all at once*. ARMSTRONG says, for example, "The logical dependence of purpose on perception and belief, and of perception and belief upon purpose is not circularity in definition. What it shows is that the corresponding concepts must be introduced *together or not at all*" (1977, p. 88). SHOEMAKER says, (1975) "On one construal of it, functionalism in the philosophy of mind is the doctrine that mental or psychological terms are in principle eliminable in a certain way." LEWIS is more explicit about this using a formulation much like the Ramsey sentence formulation given above, which designates

⁶ Characterizations of mental states along the lines of the Ramsey sentence formulation presented above wear their incompatibility with behaviorism on their sleeves in that they involve explicit quantification over mental states. Both Thomas and Bealer provide ways of transforming functionalist definitions or identifications so as to disguise such transparent incompatibility.

mental states by expressions that do not contain any mental terminology. (See his *Psychological and Theoretical Identifications*, 1972, for details.)

The same sort of point applies to machine functionalism. PUTNAM says "The S_i , to repeat, are specified only *implicitly* by the description". (1967). In the coke machine automaton described above, the only antecedently understood terms (other than 'emit', 'go to', etc.) are the input and output terms, 'nickel', 'dime', and 'coke'. The state terms ' S_1 ' and ' S_2 ' in the coke machine automaton—as in every finite automaton—are given their content entirely in terms of input and output terms (+ logical terms).

Thus functionalism could be said to reduce mentality to input-output structures (note that S_1 and S_2 can have any natures at all, so long as these natures connect them to one another and to the acceptance of nickels and dimes, and disbursement of nickels and cokes as described in the machine table). But functionalism gives us reduction without elimination. Functionalism is not fictionalist about mentality, for each of the functionalist ways of characterizing mental states in terms of inputs and outputs commits itself to the existence of mental states by the use of quantification over mental states, or some equivalent device.⁷

The varieties of functionalism

Thus far, I have characterized functionalism without adverting to any of the confusing disagreements among functionalists. I believe that my characterization is correct, but its application to the writings of some functionalists is not immediately apparent. Indeed, the functionalist literature (or rather, what is generally, and I think correctly, regarded as the functionalist literature) exhibits some bizarre disagreements, the most surprising of which has to do with the relation between functionalism and physicalism. Some philosophers (ARMSTRONG, 1968, 1970, 1977; LEWIS, 1966, 1972, 1979; SMART, 1971) take functionalism as showing that physicalism is probably *true*, while others (PUTNAM, 1966; FODOR, 1965; BLOCK and FODOR, 1972) take functionalism as showing that physicalism is probably *false*. This is the most noticeable difference among functionalist writings. I shall argue that the Lewis–Armstrong–Smart camp is mistaken in holding that functionalism supports an interesting version of physicalism, and furthermore, that the functionalist insight

⁷ The machine table states of a finite automaton can be defined explicitly in terms of inputs and outputs by a Ramsey sentence method, or by the method described in THOMAS (1978). Both of these methods involve one or another sort of commitment to the existence of the machine table states.

that they share with the Putnam–Fodor–Harman-camp *does* have the consequence that physicalism is probably false. I shall begin with a brief historical sketch.

While functionalism dates back to Aristotle, in its current form it has two main contemporary sources. (A third source, Sellars', and later, Harmans' views on meaning as conceptual role, has also been influential.)

Source I. PUTNAM (1960) compared the mental states of a person with the machine table states of a Turing machine. He then rejected any identification of mental states with machine table states, but in a series of articles over the years he moved closer to such an identification, a pattern culminating in PUTNAM (1967). In this article, Putnam came close to advocating a view—which he defended in his philosophy of mind lectures in the late 1960s—that mental states can be identified with machine table states, or rather disjunctions of machine table states. (See THOMAS, 1978, for a defence of roughly this view; and see BLOCK and FODOR, 1972, and PUTNAM, 1975, for critiques of such views.)

FODOR (1965, 1968b) developed a similar view (though it was not couched in terms of Turing machines) in the context of a functional analysis view of psychological explanation (see CUMMINS, 1975). Putnam's and Fodor's position was characterized in part by its opposition to physicalism, the view that each *type* of mental state is a physical state.⁸ Their argument is at its clearest with regard to the simple version of Turing machine functionalism described above, the view that pain, for instance, is a machine table state. Ask yourself what physical state could be in common to all and only realizations of S_1 of the Coke machine automaton described above? The Coke machine could be made of an enormous variety of materials, and operate via an enormous variety of mechanisms; it could even be a “scattered object”, with parts all over the world, communicating by radio. If someone suggests a putative physical state in common to all and only realizations of S_1 , it is a simple matter

⁸ “Physical state” could be spelled out for these purposes as: the state of something's having a first-order property that is expressible by a predicate of a true physical theory. Of course, this analysis requires some means of characterizing physical theory. A first-order property is one whose definition does not require quantification over properties. A second-order property is one whose definition requires quantification over first-order properties (but not other properties). The physicalist doctrine that functionalists argue against is the doctrine that mental properties are *first-order* physical properties. Functionalists need not deny that mental properties are second-order physical properties (in various senses of that phrase).

to dream up a nomologically possible machine that satisfies the machine table but does not have the designated physical state. Of course, it is one thing to *say* this and another thing to prove it, but the claim has such overwhelming *prima facie* plausibility that the burden of proof is on the critic to come up with reason for thinking otherwise. Published critiques (KALKE, 1969; KIM, 1972; GENDRON, 1971; CAUSEY, 1977; NELSON, 1976) have in my view failed to meet this challenge.

If we could formulate a machine table for a human, it would be absurd to identify any of the machine table states with a type of *brain* state, since presumably all manner of brainless machines could be described by that table as well. So if pain is a machine table state, it is not a brain state.

It should be mentioned, however, that it is possible to *specify* a sense in which a functional state *F* can be said to be *physical*. For example, *F* might be said to be physical if every system that in fact has *F* is a physical object, or alternatively, if every realization of *F* (that is, every state that plays the causal role specified by *F*) is a physical state. Of course, the doctrine of "physicalism" engendered by such stipulations should not be confused with the version of physicalism that functionalists have argued against (see footnote 8).

KIM (1972) objects that "the less the physical basis of the nervous system of some organisms resembles ours, the less temptation there will be for ascribing to them sensations or other phenomenal events". But his examples depend crucially on considering creatures whose functional organization is much more primitive than ours. He also points out that "the mere fact that the physical bases of two nervous systems are different in material composition or physical organization, with respect to a certain scheme of classification does not entail that they cannot be in the same physical state with respect to a different scheme". But the functionalist does not (or, better, should not) claim that functionalism *entails* the falsity of physicalism, but only that the burden of proof is on the physicalist. KIM and LEWIS (1969) (see also CAUSEY, 1977, p. 159) propose species-specific identities: pain is one brain state in dogs and another in people. As should be clear from this paper, however, this move sidesteps the main metaphysical question: "What is in common to the pains of dogs and people (and all other pains) in virtue of which they are pains?"

Source II. The second major strand in current functionalism descends from J. J. C. SMART's early article on mind-body identity (1959). Smart worried about the following objection to mind-body identity: So what

if pain is a physical state? Still, pain can have a variety of phenomenal *properties*, such as sharpness, and these phenomenal properties may be irreducibly mental. Then Smart and other identity theorists would be stuck with a “double aspect” theory: pain is a physical state, but it has both physical and irreducibly mental properties. He attempted to dispel this worry by analyzing mental concepts in a way that did not carry with it any commitment to the mental or physical status of the concepts.⁹ These “topic-neutral analyses” as he called them, specified mental states in terms of the stimuli that caused them (and the behavior that they caused, though Smart was less explicit about this). His analysis of first-person sensation avowals were of the form, “There is something going on in me which is like what goes on when...”, where the dots are filled in by descriptions of typical stimulus situations. In these analyses, Smart broke decisively with behaviorism in insisting that mental states were real things with causal efficacy; and Armstrong, and Lewis, and others later improved his analyses, making explicit the behavioral effects clauses, and including mental causes and effects. Lewis’ formulation, especially, is now very widely accepted among Smart’s and Armstrong’s adherents. (SMART, 1971, also accepts it.) In a recent review in the Australasian Journal of Philosophy, Alan REEVES (1978) declares, “I think that there is some consensus among Australian materialists that Lewis has provided an exact statement of their viewpoint”.

Smart used his topic-neutral analyses only to defeat an a priori objection to the identity theory. As far as an argument *for* the identity theory went, he relied on considerations of simplicity. It was absurd, he thought, to

⁹ As KIM has pointed out (1972), Smart did not need these analyses to avoid “double aspect” theories. Rather, a device Smart introduces elsewhere in the same paper will serve the purpose. Smart raises the objection that if after-images are brain states, then since an after-image can be orange, the identity theorist would have to conclude that a brain state can be orange. He replies by saying that the identity theorist need only identify the *experience of having an orange after-image* with a brain-state; this state is not orange, and so no orange brain states need exist. Images, says Smart, are not really mental entities: it is experiences of images that are the real mental entities. In a similar manner (notes KIM), the identity theorist can “bring” the phenomenal properties into the mental states themselves; e.g., the identity theorist can concern himself with states such as John’s having a sharp pain; this state is not sharp, and so the identity theorist is not committed to sharp brain states. This technique does the trick, though of course it commits its perpetrators to the unfortunate doctrine that pains do not exist, or at least that they are not mental entities: rather, it is the having of sharp pains and the like that are the real mental entities.

suppose that there should be a perfect correlation between mental states and brain states, and yet that the states could be non-identical. (See KIM, 1966; BRANDT and KIM, 1967, for an argument against SMART; and BLOCK, 1971, 1978; and CAUSEY, 1972, 1977, for arguments against KIM and BRANDT.) But David Lewis and Smart's Australian allies (notably D. M. Armstrong) went beyond Smart, arguing that something like topic-neutral analyses could be used to argue *for* mind-brain identity. In its most persuasive version (Lewis's), the argument for physicalism is that pain can be seen (by conceptual analysis) to be the occupant of causal role R ; a certain neural state will be found to be the occupant of causal role R ; thus it follows that pain = that neural state. Functionalism comes in by way of showing that the meaning of 'pain' is the same as a certain definite description that spells out causal role R .

Lewis and Armstrong argue from functionalism to the truth of physicalism, because they have a "functional specification" version of functionalism. Pain is a functionally specified state, perhaps a functionally specified brain state, according to them. Putnam and Fodor argue from functionalism to the falsity of physicalism, because they say there are functional states (or functional properties), and that mental states (or properties) are identical to these functional states. No functional state is likely to be a physical state.

The difference between a functional state identity claim and a functional specification claim can be made clearer as follows. Recall that the functional state identity claim can be put thus:

$$\text{pain} = \lambda y \exists x_1 \dots \exists x_n [T(x_1 \dots x_n) \& y \text{ has } x_1];$$

where x_1 is the variable that replaced 'pain'. A functional specification view¹⁰ could be stated as follows:

$$\text{pain} = \exists x_1 \exists x_2 \dots \exists x_n T(x_1 \dots x_n).$$

In terms of the example mentioned above, the functional state identity theorist would identify pain with the property one has when one is in a state that is caused by pin pricks and causes loud noises and also something else which causes brow wrinkling. The functional specifier would define pain as *the thing* that is caused by pin pricks and causes loud noises and also something else which causes brow wrinkling.

According to the functional specifier, the thing that has causal role R

¹⁰ The functional specification view I give here is a much simplified version of Lewis' formulation. See LEWIS (1972).

(e.g., the thing that is caused by pin pricks and causes something else and so forth) might be a state of one physical type in one case and a state of another physical type in another case. The functional state identity theorist is free to accept this claim as well, but what he insists on is that *pain* is not identical to a physical state. What pains have in common in virtue of which they are pains is causal role *R*, not any physical property.

In terms of the carburetor example, functional state identity theorists say that being a carburetor = being what mixes gas and air and sends the mixture to something else, which, in turn.... Functional specifiers, on the other hand, say that the carburetor is *the thing* that mixes gas and air and sends the mixture to something else, which in turn.... What the difference comes to is that the functional specifier says that the carburetor is a type of physical object, though perhaps one type of physical object in a Mercedes and another type of physical object in a Ford. The functional state identity theorist can agree with this, but he insists that *what it is to be a carburetor* is to have a certain functional role, not a certain physical structure.

At this point, it may seem to the reader that the odd disagreement about whether functionalism justifies physicalism or the negation of physicalism owes simply to ambiguities in "functionalism" and "physicalism". In particular, it may seem that the functional specification view justifies *token* physicalism (the doctrine that every particular pain is a physical state token), while the functional state identity view justifies the negation of *type* physicalism (the doctrine that *pain* is a type of physical state).

This response oversimplifies matters greatly however. First, it is textually mistaken, since those functional specifiers who see the distinction between type and token materialism clearly have type materialism in mind. For example, Lewis says, "A dozen years or so ago, D. M. Armstrong and I (independently) proposed a materialist theory of mind that joins claims of *type-type* psychophysical identity with a behaviorist or functionalist way of characterizing mental states such as pain" (LEWIS, 1979, emphasis added). More important, the functional specification doctrine *commits* its proponents to a functional identity claim. And since the latter doctrine counts against type physicalism, so does the former. It is easy to see that the functional specification view commits its proponents to a functional state identity claim. According to functional specifiers, it is a conceptual truth that pain is the state with causal role *R*. But then *what it is to be a pain* is to have causal role *R*. Thus the functional specifiers are com-

mitted to the view that what pains have in common in virtue of which they are pains is their causal role, rather than their physical nature. (Again, LEWIS, 1979, is fairly clear about this: "Our view is that the concept of pain... is the concept of a state that occupies a certain causal role.")

I suspect that what has gone wrong in the case of *many* functional specifiers is simply failure to appreciate the distinction between type and token for mental states. If pain in Martians is one physical state, pain in humans another, and so on for pain in every pain-feeling organism, then each particular pain is a token of some physical type. This is token physicalism. Perhaps functional specifiers ought to be *construed* as arguing for token physicalism (even though Lewis and others explicitly say they are arguing for type physicalism). I shall give three arguments against such a construal. First, as functional state identity theorists have often pointed out, a *nonphysical* state could conceivably have a causal role typical of a mental state. In functional specification terms, there might be a creature in which pain is a functionally specified *soul* state. So functionalism opens up the possibility that even if *our* pains are physical, other pains might not be. In the light of this point, it seems that the support that functionalism gives even to token physicalism is equivocal. Second, the *major* arguments for token physicalism involve no functionalism at all (see DAVIDSON, 1970, and FODOR, 1974). Third, token physicalism is a much weaker doctrine than what physicalists have typically wanted.

In sum, functional specifiers *say* that functionalism supports physicalism, but they are committed to a functionalist answer, not a physicalist answer, to the question of what all pains have in common in virtue of which they are pains. And if what all pains have in common in virtue of which they are pains is a functional property, it is very unlikely that pain is coextensive with any physical state. On the other hand, if functional specifiers have *token* physicalism in mind, functionalism provides at best equivocal support for the doctrine; better support is available elsewhere, and the doctrine is a rather weak form of physicalism to boot.

Lewis's views deserve separate treatment. He insists that pain is a brain state only because he takes 'pain' to be a non-rigid designator meaning 'the state with such and such causal role'.¹¹ Thus on Lewis' view, to say that pain is a brain state should not be seen as saying what all pains have

¹¹ A rigid designator is a singular term that names the same thing in each possible world. 'The color of the sky' is non-rigid since it names blue in worlds where the sky is blue, and red in worlds where the sky is red. 'Blue' is rigid, since it names blue even in worlds where the sky is red.

in common in virtue of which they are pains—just as saying that the winning number is 37 does not suggest that being 37 is what all winning numbers have in common. Many of Lewis' opponents disagree about the rigidity of 'pain', but the dispute is irrelevant to our purposes, since Lewis does take 'having pain' to be rigid, and so he does accept (he tells me) a functional property identity view: having pain = having a state with such and such a typical causal role. I think that most functional state identity theorists would be as willing to rest on the thesis that having pain is a functional property as on the thesis that pain is a functional state.

In conclusion, while there is considerable disagreement among the philosophers whom I have classified as metaphysical functionalists, there is a single insight about the nature of the mind to which they are all committed.¹²

References

- ARMSTRONG, D. M., 1968, *A materialist theory of mind* (Routledge & Kegan Paul, London)
- ARMSTRONG, D. M., 1970, *The nature of mind*, in: *The Mind/Brain Identity Theory*, C. V. Borst, ed., (Macmillan, London)
- ARMSTRONG, D. M., 1977, *The causal theory of the mind*, in: *Neue Heft für Philosophie*, no. 11, pp. 82–95 (Vandenhoek und Ruprecht)
- BEALER, G., 1978, *An inconsistency in functionalism*, *Synthese*, vol. 38, pp. 333–372
- BLOCK, N., 1971, *Physicalism and theoretical identity*, Ph. D. dissertation, Harvard University
- BLOCK, N., 1978, *Troubles with functionalism*, in: *Minnesota Studies in Philosophy of Science*, vol. 9, ed. C. W. Savage
- BLOCK, N., 1979, *Readings in philosophy of psychology* (Harvard University Press, Cambridge)
- BLOCK, N., *Are absent qualia impossible?*, to appear in *The Philosophical Review*, forthcoming
- BLOCK, N., and J. A. FODOR, 1972, *What psychological states are not*, *Philosophical Review*, vol. 81, no. 2, pp. 159–182
- BRANDT, R., and J. KIM, 1967, *The logic of the identity theory*, *Journal of Philosophy*, vol. 64, no. 17, pp. 515–537
- CAUSEY, R., 1972, *Attribute identities in micro-reductions*, *Journal of Philosophy*, vol. 69, no. 14, pp. 407–422
- CAUSEY, R., 1977, *Unity of science* (Reidel, Dordrecht)
- CHISHOLM, R. M., 1957, *Perceiving* (Cornell University Press, Ithaca)
- CUMMINS, R., 1975, *Functional analysis*, *Journal of Philosophy*, vol. 72, no. 20, pp. 741–764
- DAVIDSON, D., 1970, *Mental events*, in: *Experience and Theory*, eds. L. Swanson and J. Foster (University of Massachusetts Press, Amherst)

¹² Earlier versions of this material appeared in the introduction to the section on functionalism in BLOCK (1979), and in BLOCK (1978) (also reprinted in BLOCK, 1979).

- DENNETT, D., 1975, *Why the law of effect won't go away*, Journal for the Theory of Social Behavior, vol. 5, pp. 169–187
- FIELD, H., 1978, *Mental representation*, Erkenntnis, vol. 13, pp. 9–61
- FODOR, J. A., 1965, *Explanations in psychology*, in: *Philosophy in America*, ed. M. Black (Routledge & Kegan Paul, London)
- FODOR, J. A., 1968a, *The appeal to tacit knowledge in psychological explanation*, Journal of Philosophy, vol. 65, pp. 627–640
- FODOR, J. A., 1968b, *Psychological explanation* (Random House, New York)
- FODOR, J. A., 1974, *Special sciences*, Synthese 28
- FODOR, J. A., 1975, *The language of thought* (Crowell, New York)
- GEACH, P., 1957, *Mental acts* (Routledge & Kegan Paul, London)
- GENDRON, B., 1971, *On the relation of neurological and psychological theories: A critique of the hardware thesis*, in: *Boston Studies in the Philosophy of Science*, eds. R. C. Buck and R. S. Cohen, vol. 8 (Reidel, Dordrecht)
- GRICE, H. P., 1975, *Method in philosophical psychology (from the Banal to the Bizarre)*, Proceedings and Addresses of the American Philosophical Association (American Philosophical Association, Newark, Del.)
- HARMAN, G., 1973, *Thought* (Princeton University Press, Princeton)
- HARTMAN, E., 1977, *Substance, body and soul* (Princeton University Press, Princeton)
- KALKE, W., 1969, *What is wrong with Fodor and Putnam's functionalism?*, Nous, vol. 3, pp. 83–93
- KIM, J., 1966, *On the psycho-physical identity theory*, American Philosophical Quarterly, vol. 3, no. 3, pp. 227–235
- KIM, J., 1972, *Phenomenal properties, psychophysical law, and the identity theory*, Monist, vol. 56, no. 2, pp. 177–192
- LEWIS, D., 1966, *An argument for the identity theory*, reprinted in: *Materialism and the Mind-Body Problem*, ed. D. Rosenthal (Prentice-Hall, Englewood Cliff, N. J., 1971)
- LEWIS, D., 1969, *Review of 'art, mind and religion'*, Journal of Philosophy, vol. 66, no. 1, pp. 23–35
- LEWIS, D., 1970, *How to define theoretical terms*, Journal of Philosophy, vol. 67, no. 1, pp. 427–444
- LEWIS, D., 1972, *Psychophysical and theoretical identification*, Australasian Journal of Philosophy, vol. 50, no. 3, pp. 249–258
- LEWIS, D., *Mad Pain and Martian Pain*, in: BLOCK, 1979
- LYCAN, W., *A new lilliputian argument against machine functionalism*, Philosophical Studies, forthcoming
- MARTIN, R. M., 1966, *On theoretical constants and Ramsey constants*, Philosophy of Science, vol. 31, pp. 1–13
- NAGEL, T., 1970, *Armstrong on the mind*, Philosophical Review, vol. 79, pp. 394–403
- NELSON, R. J., 1975, *Behaviorism, finite automata and stimulus response theory*, Theory and Decision, vol. 6, pp. 249–267
- NELSON, R. J., 1976, *Mechanism, functionalism and the identity theory*, Journal of Philosophy, vol. 73, no. 13, pp. 365–386
- PUTNAM, H., 1960, *Minds and machines*, in: *Dimensions of Mind*, ed. S. Hook (New York University Press, New York)

- PUTNAM, H., 1963, *Brains and behavior*, reprinted in: Mind, Language, and Reality: Philosophical Papers, vol. 2 (Cambridge University Press, London 1975)
- PUTNAM, H., 1966, *The mental life of some machines*, reprinted in: Mind, Language and Reality: Philosophical Papers, vol. 2 (Cambridge University Press, London 1975)
- PUTNAM, H., 1967, *The nature of mental states* (originally published as Psychological Predicates), reprinted in: Mind, Language and Reality: Philosophical Papers, vol. 2 (Cambridge University Press, London 1975)
- PUTNAM, H., 1970, *On properties*, in: Mathematics, Matter and Method: Philosophical Papers, vol. 1 (Cambridge University Press, London)
- PUTNAM, H., 1975, *Philosophy and our mental life*, reprinted in: Mind, Language and Reality: Philosophical Papers, vol. 2 (Cambridge University Press, London)
- REEVES, A., 1978, *Review of W. Matson "Sentience"*, Australasian Journal of Philosophy, vol. 56, no. 2 (August), pp. 189–192
- REY, G., 1979, *Functionalism and the emotions*, in: Explaining Emotions, ed. A. Rorty, (University of California Press, Berkeley and Los Angeles)
- RYLE, G., 1949, *The concept of mind* (Hutchinson, London)
- SELLARS, W., 1968, *Science and metaphysics* (Routledge & Kegan Paul, London), chap. 6
- SHOEMAKER, S., 1975, *Functionalism and qualia*, Philosophical Studies, vol. 27, pp. 271–315
- SMART, J. J. C., 1959, *Sensations and brain processes*, Philosophical Review, vol. 68, pp. 141–156
- SMART, J. J. C., 1971, *Reports of immediate experience*, Synthese, vol. 22, pp. 346–359
- THOMAS, S., 1978, *The formal mechanics of mind* (Cornell University Press, Ithaca)

THE IMPLICATIONS OF LAND'S THEORY OF COLOUR VISION

KEITH CAMPBELL

University of Sydney, Sydney, Australia

1. The situation prior to Land's contribution

There are two critical areas of difficulty for any metaphysical materialism in the philosophy of mind; the interpretation of intentionality, and the reduction of secondary qualities. In the first instance at least, a theory of colour vision is, of course, a contribution to the debate over secondary qualities. Dr. E. H. Land, of the Polaroid Corporation and instant photography, has made proposals which amount to a new theory of the whole process of seeing in colour.¹ His accomplishment can be summarized, I believe, very briefly: hitherto, the colours have been in a worse position, so far as the prospects of a satisfactory reduction are concerned, than other secondary qualities. If Land is fundamentally right, colours are in the same condition, no worse (and no better) than sounds, smells, and perceived warmth.

Colours were formerly in a worse position than, for example, sounds, because strenuous efforts had failed to find a single physical reality corresponding to each of the different colours. By great misfortune, investigations of the physical basis of colour began with the celebrated Newtonian prismatic resolution of sunlight into its spectrum. With the development of the wave theory of light, the association of distinctive spectral colours with their own distinctive wavelength was inevitable.

The subsequent discovery of emission spectra for the different elements,

¹ An account accessible to laymen is to be found in *Scientific American*, vol. 237, no. 6, Dec. 1977.

consisting of narrow bands of emission at particular wavelengths, each of a characteristic hue, further consolidated the view that colour is, more or less directly, a matter of wavelength.

Indeed, this view became so entrenched that many philosophers still believe it. It is, however, false.

If the light coming from any surface is concentrated in a narrow waveband, then that surface will indeed have one distinctive colour under a wide variety of observational circumstances. Such surfaces are, however, very exceptional. The great majority of coloured surfaces are selective reflectors of incident light. They reflect to some degree across at least a considerable segment of the visible spectrum. The great majority of illuminants providing the incident light contain at least some light at all visible wavelengths. Selective reflection of such light sends to the eye light with a component at most, if not all, the wavelengths to which the eye is sensitive.

Ordinarily, two surfaces which look different in colour send to the eye light containing different proportions of the various visible wavelengths. The *flux* at a given wavelength is the amount of energy (the intensity) in that wavelength's light. Total energy across a range of wavelengths can be arrived at by integration. Different amounts of energy at the various wavelengths will yield different flux profiles for light of different compositions. Under conditions in which two surfaces receiving the same incident illumination look different in colour, they will be found to be sending different fluxes to the observer. Hence flux became a natural candidate for the physical basis of colour.

One great stumbling block to flux theories of colour lay in the discovery that not only do surfaces of differing colours deliver different fluxes to the observer, but so do many surfaces between which normal human colour vision does not distinguish.

Even if we put to one side anomalous or unusual cases, such as the colours produced by spinning a white disc with a black spiral on it, or colours in after images, or colours seen on a thin oil film, there remains a wide range of fluxes, emanating from normal selective reflectors, associated with any given specific colour. Worse, no unifying formula could be found to fit all and only these various fluxes belonging to one specific colour. The attempt to identify flux as the physical correlate of colour did not succeed. The search for a physical basis in flux proves to be a long wavelength herring.

With hindsight, we should have suspected this. The colours we see

objects as having do not vary even under quite a wide range of variations in the local illumination, from indoor to outdoor, or noon to evening, or sunlight to overcast, which alter the flux leaving their surfaces.

2. Land's negative achievement

A major part of Land's contribution to the problem lies in a series of experiments demonstrating the independence of colour and flux, and so explaining why no unifying flux formulae for the different colours are forthcoming.

Land set up experiments in which the flux reaching the eye could be known with precision. He did this by using as illuminant a set of three narrow wave-band projectors, one each in the long, middle, and short wavelength ranges of the visible spectrum (a red, a green, and a blue).

Many sets of three such light sources, combined in varying proportions, and sometimes diluted with white, can replicate any colour discernible by normal humans—such sets of lights are known as sets of *additive primaries*. Colour television sets reproduce colour by a mosaic array of additive primaries.

Since a set of additive primaries matches the whole visible spectrum in the richness of its colour producing powers, use of such a set in place of whole-spectrum illuminants seems an acceptable simplification.

In Land's experiments, the three illuminators light up large square arrays of patches of many different colours, arranged so that each patch is surrounded by several differently coloured ones. The arrays are nicknamed 'colour Mondrians': two identical colour Mondrians are illuminated by two identical sets of narrow band projectors.

On one colour Mondrian a white patch is selected. The intensity of light it reflects to the eyes at each of the three incident wavelengths (the flux) is measured. On the second colour Mondrian another patch, say a green one, is chosen. The intensity of light from the second set of projectors is now adjusted so that the flux from the green patch matches, wavelength for wavelength, that from the original white patch.

Under these circumstances, despite identity of flux, the white looks white and the green looks green. The two patches are viewed simultaneously, so no rapid adaptation phenomenon nor any trick of memory is involved.

Moreover, the projectors can be further adjusted so that a red or a blue patch, keeping its red or blue appearance, can send the same flux to the eye as the original white and green. Land goes so far as to claim that

any colour on his Mondrian can be made to send the original flux to the observer.

In short, the same light, entering eyes in the same condition, can give rise to impressions of colour from seemingly anywhere on the colour wheel. One could scarcely ask for a more convincing display of the independence of colour and flux.

One could, however, ask for some indication of what, if not flux, constitutes the physical basis of colour. For colour vision is not capricious. Non-collusive agreement over colour qualities and colour changes is copious and subtle. Inter-subjective co-incidence in judgment calls for an objective reality underpinning it.

3. Land's positive experiments

In a second series of experiments Land takes up the positive task of identifying the real physical correlate for colour. Here simplified colour Mondrians, with only 17 colour patches, are used in conjunction with the *Munsell Book of Colour*, which contains over 1.000 standard colour 'chips' for use in matching tests. The strategy is admirable: take a group of colour Mondrian patches, and adjust the illuminators till each sends the same flux to the eye. Match each in turn with a Munsell chip, which, although indistinguishable in colour from one of the Mondrian patches, will in fact be sending the eye a quite different flux.

The original negative conclusion, that flux is neither common to all surfaces of the same colour, nor peculiar to surfaces of that colour only, is reinforced.

The matching pairs, Mondrian patch and Munsell chip, are the objects of study. Some physical feature which both share, but which no other non-matching surface possesses, will be a good candidate for the physical basis we seek, common and peculiar to surfaces of each specific colour considered in turn.

4. The theory built on these experiments

At any given wavelength, a surface will reflect some incident light and absorb some. The proportion reflected at a given wavelength is the *reflectance* at that wavelength. Coloured surfaces, as contrasted with white, grey or black ones, have different reflectances at different points on the spectrum, which is why they are described as selective reflectors.

In the Mondrian-Munsell matching experiments, both the intensity of incident illumination and intensity of reflected light are known for each of the three projected wavelengths. From this the reflectance at those wavelengths is determined.

Land is able to show that the reflectances of matching surfaces match, while the reflectances of differently coloured surfaces do not match.

To be precise, he shows this for the three wavelengths at which his projectors illuminate the scene. It requires a further step to check that reflectances match or are otherwise equivalent right across the spectrum. So far as I know, this further step has not yet been taken. But there is at this point no reason to think that it would prove recalcitrant.

A match in reflectance is a match in how a surface modifies light in reflecting it. This modification is plainly a different matter from its product, the composition of the light coming to an observer. A match in reflectance is different from a match in consequent flux.

Provided they receive the same illumination, surfaces with the same reflectance will of course send to the eye the same flux. When experimenters controlled illumination, they created the situation in which matching reflectance coincides with matching flux at the eye. Controlling the illumination in experiments on colour vision, an apparently obvious *desideratum*, thus ironically misled by appearing to cement a link between colour and flux.

Constant illumination not only gives flux a misleading constancy. It diverts attention from the striking phenomenon of colour constancy under varying illumination.

Under varying illumination reflected flux varies but reflectance does not. Within wide limits, nor does colour. This is the key fact upon which Land proposes to build.

5. The mechanisms involved in colour vision

Apart from the supersensitive rods, which generate the monochrome field of various greys with which we are familiar from experience of the world in twilight, there are in the eye three cone systems each responding over a part of the visible spectrum, each reaching peak sensitivity at a different wavelength. There is no doubt a connection between the existence of three cone systems and the need for exactly three well-chosen illuminators to serve as additive primaries for normal ('trichromatic') perceivers.

According to Land, the primary determination which visual receptor systems make concerns how light a surface is. Lightness is a familiar quantity; white surfaces have it in high degree, various greys in diminishing amount, and blacks practically not at all. Lightness is a feature belonging to appearance; lightness values are established by getting observers to judge a surface's position relative to black and white. Such judgments correspond closely to the relative reflectance of the surface in question. Although we cannot isolate and stimulate cone systems individually, Land holds that each of the cone systems makes its own independent judgment of lightness. Each surface we see is, so to speak, seen three times over, on three different, though overlapping, wavebands. And on each of those wavebands, the human visual system establishes how light the surface is. The result is a triple of lightness judgments over long-, middle-, and short-wave segments of the visible spectrum. These lightness values are the appearances to observers of reflectance values in the surface which is being seen.

Each surface has a reflectance at every wavelength, and this can be integrated across the range to which each cone system is sensitive. This yields a triplet of reflectances, which constitute the physical basis of colour. The reflectances give rise to a triplet of lightnesses, which constitute the observational basis of colour. To every discernible colour corresponds a unique triplet of lightnesses. That is the essence of Land's doctrine.

Three lightness values means three dimensions of variation, so the colours can all be located on a Colour Cube, with each colour's unique lightness triple serving as a uniquely locating set of coordinates. Black, darkest on all three cone systems, is at the origin, white at the opposite vertex, the greys along the diagonal representing equal lightness on each system. The other colours are dispersed through the space of the cube.

6. Determining reflectance triples

But now the problem to be faced is this: the information on which colour vision works must all be contained in the light reaching the observer. However reflectance, and hence how light things look, consists in the relationship between illumination and reflected flux. How can reflectance be judged without knowledge of the composition and strength of the illumination?

Land's approach to this problem is by way of *comparison* of flux coming from different surfaces at the same time. Hence his use of polychrome

colour Mondrians. Take one cone system, say the short-wave sensitive one. Under illumination uniform in the range to which that system responds, different surfaces in the visual field will send different fluxes to an observer. This will enable a rank order of lightnesses to be established. The absolute levels of flux from lightest and darkest areas, and hence the range of lightness values, will also be available.

Where the range of lightnesses is wide, the lightest surface will be highly reflective over the short-wave band, and the ratios other, darker, surfaces bear to the lightest will be closely related to their reflectances.

I think less favourable cases will involve comparisons not only within one cone system, but between different systems.

A narrow range of lightness at flux values which are rather high in the context of the total illumination can be produced in two ways; either by a set of surfaces all of which reflect strongly at the short end of the spectrum, or by an illumination whose composition is skewed towards the short end. The result in either case is that everything has a bluish cast. The effect can be produced in the first way by painting everything blue, or in the second way by using a blue plastic lamp shade.

Comparisons of flux values between different cone systems can likewise establish lightness values for a narrow range all located at the darker end.

Thus lightness triples can be established for the three cone systems, and by their means we can get to reflectance triples, and so to colours.

7. The virtues of this theory

If Land is right, every variation in colour can be correlated with a specific physical variation. In principle, the colour which a surface will have in any specified illumination can be calculated and predicted in advance. Colour will be in the same position as sound, or felt temperature, or, if stereo-chemical theory is on the right track, smell.

So far as I can see, Land's doctrine can accommodate most of the colour variation phenomena, such as change of colour on very close approach, or under a microscope, or in the distance, or through haze, or tinted glass. These are all cases where a non-standard relationship between flux leaving a surface and flux entering the eye can affect judgment on flux and hence judgment on reflectance. With objects in shadow, the darker shift arises from non-standard illumination.

Every theory of colour needs a comfortable machinery to accommodate our dual way of ascribing colours. On one use colours are standing

phenomena, changed only by intrinsic change in the object, such as a leaf undergoes in autumn or a motor car at the spraypainters. Standing colours remain the same through changes in illumination or other extrinsic circumstances. Hills do not change from blue to green on closer approach, nor does water go pink in the dawn.

On another use, colour terms pick out just the occurrent hue of the present moment, and coloured spotlights in the theatre are said to transform the colours of skin, clothing and sets.

There are two correct answers to Locke's question about porphyry in the dark, one for the standing, one for the transitory colour.

Land's theory provides the required machinery for dealing with this double use. In the object, *reflectance* triples provide an objective basis for an intrinsic standing colour—just which one will be determined by the apparent colour in standard noon-day conditions.

In experience, the *lightness* triples are subject to variation even when reflectance is not altered, by change in distance, illumination, and state of observer. Lightness triples provide the basis for transitory colours.

The theory is attractive in another way too. Colours apparently play a negligible role in causal chains in the inanimate realm. Notoriously, in offering physical explanations we have little occasion to invoke the colours of objects. In Stout's phrase, colours seem not to belong to the executive order of nature. This has been one source of subjectivism about colours.

But here we have the materials for an objectivist account of the situation: if colours are integrated reflectances across three overlapping segments clustered in the middle of the total electromagnetic spectrum, then they are, from the inanimate point of view, such highly arbitrary and idiosyncratic properties that it is no wonder the particular colours we are familiar with are manifest only in transactions with humans, rhesus monkeys, and machines specially built to replicate just their particular mode of sensitivity to photons.

Another virtue of the theory is its properly empirical character. It claims that a certain quite specific mode of processing flux inputs at the eye is performed somewhere in the retina-optic nerve-cortex complex. Just where is not yet specified—hence Land's use of the term 'retinex' to cover the proposed location of the processing activity.

A computer programme has made good progress towards reproducing the flux-comparison processing which the theory requires. If there are neural functions which constitute the integrating and comparing activities

involved, it is, in principle, possible to identify them. If on the other hand the nervous system is not capable of performing the required tasks, it is in principle possible to establish that. This is as it should be.

In this connection, it is noteworthy that independent work by Dr. A. L. Gilchrist confirms the importance to vision of contrasts at edges in the visual field.² These are crucial to interpretation, and can enable judgments about illumination to be made, which is plainly a bonus in the business of determining reflectance.

8. Its problems

There are, however, some unresolved problems in the theory. As it stands, it is organised to account only for coloured *reflecting* surfaces. It needs further elaboration to cope with visual fields containing both emitters and reflectors. For, of course, a light source has a colour, and so presumably a set of lightness values, but these lightness values are not determined by its reflectance at all. Is a light source identified, perhaps, by its exceptionally high flux values, then treated by considering relativities between the three cone systems as if they were produced by reflection under white light?

Further, there are difficulties with visual fields which are not variegated and polychromatic. Land discusses the spot-in-the-void, a single small source of narrow band light on a black ground. How can its lightness triple be determined, where no comparisons are available for fixing relative lightnesses? According to Land, comparisons between the three cone systems suffice. The three systems respond differently, and the relationship between these responses is nearly invariant with changes in intensity of the light source.

But the cone systems are responding to flux, and lightness is supposed to relate not to flux but to reflectance. The same relative responses, on the three systems, can be produced by many different reflectances under different illuminations. If the spot were not an emitter but a reflector, how could we tell that it is a very narrowly selective reflector without knowledge of the illumination? A *reflecting* spot in a void does not provide enough information for a Land-type determination of colour. I see no basis on which arriving at the lightness triple for emitters of light should be significantly easier.

² Scientific American, vol. 240, no. 3, March 1979, pp. 88ff.

The position is not better, indeed is potentially worse, with a monochrome field sending light of all wavelengths to the eye, such as a cloudless summer sky viewed while lying on your back, or a colour card brought right up to your nose. I do not know how lightnesses can be established in these conditions.

9. Prospects for a materialist reduction of colour vision

Let us accept that Land's work lays at least the foundations for a satisfactory theory of colour vision. Do colours now cease to be an embarrassment for materialist theories of the mind?

No, they do not. As already mentioned, colours move into a position comparable to the other secondary qualities, sounds, smells, tastes, felt warmths. But from a materialist perspective, that is not a particularly satisfactory position.

We must here face once more the old issue of *qualia*. It seems such an arbitrary and contingent matter that sundry lightness triples should appear as scarlet, yellow or ultramarine. To which the natural reply is: well the lightness triples must look like *something*, why shouldn't they be seen as colours? Why shouldn't the colours be no more than ways in which lightness triples are seen, hence intentional characters no more troublesome to materialism than intentionality itself?

I cannot regard that as a satisfactory reply. Contrast the case of colour with that of the true-blue primary qualities. Take position, or shape, or size, or orientation. Each of these physical characters is perceived, visually, through the occurrence of position, shape, size, or orientation in the visual field. The perceptual character through which each of these primary qualities is perceived is that very quality itself. Likewise for changes in these primary qualities. Change of position—movement—is perceived through movement in the visual field. The growth, deformation, and rotation of bodies can be seen thanks to growth, deformation, and rotation in the visual field. There is no gulf between the physical fact and its appearance.

But how different are the colours! What natural link could lead us to expect particular lightness triples to present themselves as emerald, indigo, or rose? Or that changes in lightness triples would be manifest as a kaleidoscope?

To highlight the distinction between colour and reflectance, consider selective fatigue. If you sit for a while in a room full of long wavelength

light, the long wavelength cone system will get fatigued relative to the other two systems, and will respond less than normal to further stimulation. If you now step outside and look around, the visual system will react as if there were a systematic decline in relative long wavelength lightness values. The effect is comparable to that of adding to daylight an additional middle- and short-wave illuminant. The scene will appear as if bathed in mixed blue and green, that is cyan auxiliary light. Colours will shift systematically. Reds will go dark. Yellows will go greenish. Magentas will go blue. Whites and greys will drift to cyan.

Now so far as we can tell, it could perfectly well have happened that humans evolved with a long-wave cone system more weakly sensitive than the others. If that had happened, standard conditions would have yielded us these experiences of colour which are now obtainable only under selective fatigue. In that case, while reflectances would have remained the same, colours would have been systematically transformed. This possibility shows that in the absence of a particular mode of sensibility which we humans have there is nothing peculiarly or intrinsically *red* about a certain reflectance triple. The connection between reflectance triples and the colour experiences to which they give rise is looser and more contingent than in the case of size, shape, or solidity, for which no comparable systematic transformation seems possible.

If colours resist identification with the appropriate physical characteristics of physical surfaces, perhaps seeing the colours can be identified with the corresponding physiological processes in the 'retinex' system?

Two sorts of considerations suggest otherwise. First, the mutual independence of the three lightness judgments over the three ranges is of the essence of Land's theory, and this makes a lightness triple analogous to a musical chord. But unlike the musical case, the experience of scarlet contains no independent elements which nature or training could enable us to distinguish as its three components.

It could be urged in reply that in the perception of colour, lightness values are held separate and compared with one another at an early processing stage, from which only a single, summary, resultant response emerges into consciousness.

This does not, in my view, dispose of the difficulty entirely. It accounts for the absence of any chord-like structure in the colours, but leaves us in the dark as to how seeing scarlet, for example, could be doing a complex set of transformations on lightness inputs which we have no reason to suppose are themselves inherently colour experiences. The experience of

seeing scarlet has none of the characteristics which experiences of abstracting, comparing, and combining typically have. And *they* on the other hand, are diaphanous. Hue is one feature they most conspicuously lack.

The second sort of ground for resisting a physiological identification for colour vision is structural.

Changes in lightness and changes in colour are not isomorphic. Equal lightness differences are by no means tied to equal colour differences. On the colour cube, adjacent places in some regions have colours much more different than equally adjacent places elsewhere. Reds and greens furnish perhaps the most striking example. A strong red and an equally vivid green may differ only in a small change in lightness in the middle wave-length range. Comparable changes in other contexts result in no more than a slight change in hue.

This is not, so far, a decisive point. It may be argued that survival adaption has built in exaggerations at certain points, so that we habitually mistake certain degrees of colour difference. Perhaps in that case the colour cube will teach us to mend our ways and lead us to admit that reds and greens are really close to one another.

It may be argued that when we sit down to test the question, we will find that we can discriminate as many different shades between two initially very alike blues, adjacent on a 'coarse-texture' colour cube, as we can discriminate intermediates between the red and the green adjacent on the coarse-texture cube. Such a result would reveal a sort of colour-distance illusion phenomenon.

I rest no confidence in these lines of reply. For to some degree colour experience is essentially diverse. The visual field gets some of its quality from being a field of colour contrasts and colour likenesses. There could be a monochrome world. But whatever its colour, it would not be *that* colour unless it would differ in colour, to just such and such a degree, from other possible ones. Reds cannot stay reds while somehow becoming, or seeming to become, more like greens.

Qualia will not go away. They belong to mental life on its experiencing side, and resist reduction. They suggest the experiencing mind's existence is in some measure independent of its material base.

INTENTIONALITY AND BEHAVIORISM

DAGFINN FØLLESDAL

Stanford University, Stanford, U.S.A., and University of Oslo, Oslo, Norway

1. Behaviorism

“Behaviorism” stands for a variety of attitudes and methodological positions in psychology, from on the one side the epistemological view that in studying man, the sole evidence that our theories can be tested against is observation of his behavior, to on the other extreme, various rather restricted ontological views concerning what man is. In between there are all kinds of positions whose definitions are often so varied and vague that not even their various proponents agree on them,¹ like logical behaviorism, philosophical behaviorism, methodological behaviorism, radical behaviorism, neo-behaviorism, etc. The practitioners are not much to be blamed for this; “-isms” of all kinds, including for example “positivism”, “existentialism”, etc. are notoriously difficult to define. As Suppes has pointed out,² we have no well-defined framework for such definitions, as we have for definitions within mathematics or physics or any well developed science. The “-isms” are as varied as their practitioners, and for this reason I find all discussion of “-isms” for polemical purposes fruitless—not only wasteful and uninteresting, but even definitely harmful, insofar as such criticism makes somebody reject off-hand viewpoints and research findings without even having studied them.

For this reason, I shall not attempt any wholesale evaluation or condemnation of behaviorism. I will only pick out certain notions that recur in many versions of behaviorism, notably “stimulus” and “response”,

¹ Thus, for example, Michael MARTIN, in *Interpreting Skinner* (1978), argues that although Skinner explicitly denies that he is a methodological behaviorist, in one sense he is one, and in one sense he is also a philosophical behaviorist.

² SUPPES (1969), p. 294. Cf. also SUPPES (1975), pp. 269–285.

in order to see how they fare and have to be reconsidered in connection with the phenomenon that is usually regarded as the main stumbling-block of behaviorism, *intentionality*.

2. Intentionality

By "intentionality", I do here not mean primarily the practical notion of intending to do something, but the Brentano–Husserl notion of the *directedness* of the mental.

Let us first make a little more precise what is meant by this. While for Brentano the directedness of the mental simply meant that for each mental phenomenon, e.g. for each case of perception, there is some object *towards which* it is directed, *of* or *about* which it is, Husserl had a more discerning view. He acknowledged that many mental phenomena, e.g. hallucinations, do not have any object. Rather than attempting to account for intentionality by appeal to an object that the mental phenomenon is directed towards, Husserl focused on what the directedness consists in: what are the features of the mental thanks to which it always is *as if* it has an object. Husserl called the collection of these features the *noema*. He regarded the noema as a generalized notion of meaning, thereby tying together intention with a *t* and intension with an *s*.

I shall not go into Husserl's particular analysis here. I will focus only on one special feature of the analysis which is of importance in what follows: According to Husserl, what we perceive in a given case of perception is under-determined by the physical stimuli that we receive. An example that illustrates this, is Jastrow's duck-rabbit example which has become so well known through Wittgenstein. Although the lines on the paper and the pattern of irradiation on our retina remain the same, we can see a duck or a rabbit. As the plethora of this kind of examples shows, the observation is not special for Husserl, it is old and universally agreed upon. What is a little more specific to Husserl, is that he holds that *all* perception is under-determined like this. Also, his view is not quite well illustrated by the duck-rabbit example since in that example there is an external physical object, the lines on paper, that remains the same and is "taken" in different ways. For Husserl, there is *not* some basic, given object that is *seen as* a duck or *as* a rabbit. In the normal case of perception we see a rabbit or a duck directly, there is no primitive object which we see and which is then *interpreted* in different ways.

Only in special cases, like the duck-rabbit case, are we aware of the

ambiguity, of our freedom to perceive one thing *or* the other. In most cases we perceive one thing, and only if later experience forces us to give up our original belief in this thing, do we discover that there is some other thing there, perhaps a quite different one.

Now, if Husserl is right about perception, as I think he is, then this has consequences for the behaviorist notion of stimulation. Behaviorists are faced with a dilemma here. Either they may describe the stimulation as an object perceived by the subject, a ring, a color, etc. Then the problem is: how does the experimenter know that this is the object perceived by the subject? Does not the experimenter here impose his own conception on the situation on the subject that he describes? Arne Naess discussed this kind of imposition in a somewhat different framework in his doctoral dissertation in 1936, and labelled it "maze epistemology" (*Labyrinthherkenntnistheorie*) (NAESS, 1936, esp. pp. 53 ff.). Ever since von UEXKÜLL's (1921 and 1928) attempts in the early part of the century to study the environments of animals as *conceived* or *lived in* by the animals, it has been a formidable challenge for the behavioral scientist *not* to impose his own conceptions on the subjects studied.

Any attempt to describe the stimuli as *that which the subject perceives* is faced with this problem, if Husserl is right.

3. Reception and perception

Trying to avoid this, the sophisticated behaviorist might get the idea of defining the stimuli not as something that the subject perceives, but as something that goes on at his nerve endings. This is what Quine does in *Word and Object* (QUINE, 1960, pp. 31–35). When Quine defines stimuli as evolving patterns of irritation of the nerve endings, he does not hold that the stimulus is something that the subject perceives, but that it is something that he receives. Quine's central concern is therefore *not* to describe the stimuli in some neutral, physical vocabulary. Rather, Quine is aware that if he were to describe stimuli as objects perceived, he would be begging the questions concerning meaning and reference that he set out to clarify. His use of the physical vocabulary is due solely to the fact that physics, notably elementary optics, is well developed and a relatively uncontroversial part of our current world scheme; so that the descriptions of the stimuli will be precise and not subject to disagreement among those who study man. Quine does *not* impute this physical theory to the agents who are studied. The stimuli are simply not what they perceive.

However, here we are at the other horn of the behaviorist dilemma. For what Quine and other behaviorists are after, are the systematic connections between the stimuli that a subject receives and his responses. The responses, however, are responses to what the subject *perceives* and not merely to what he *receives*. Whether a subject assents to or dissents from "Gavagai", for example, depends on whether he sees a duck or a rabbit. As we have noted, the stimulus may be the same, but the objects perceived widely different—and so the responses.

Quine is aware of this. In his latest book, *The Roots of Reference*, he introduces the distinction between what a subject receives and what he perceives. The crucial difficulty in seeking to clarify perception on the basis of what is received and overt behavior is, as we should expect, to make proper allowance for interferences from within, from what speaking mentalistically we could call the subject's mental states and processes. Quine does not want to screen out the subject's actual contributions to perception, but ends up with the view that "mental entities are unobjectionable if conceived as hypothetical physical mechanisms and posited with a view strictly to the systematizing of physical phenomena" (QUINE, 1974, pp. 33–34). Hence, at least in the case of one prominent behaviorist, the difficulties connected with the notion of what is perceived have led to the admission of mental entities as part of our theory of man. Whether these mental entities are also physical, or whether they are merely mental, is, it seems, a matter that cannot be decided *a priori*.

4. Digression: Suppes' behavioristic definition of "sign"

Patrick Suppes has argued (SUPPES, 1969, pp. 300 ff.) that the difficulties connected with accounting for intentional notions, like e.g. the sign relation, within a behavioristic framework, are not as formidable as they might seem. Chisholm has claimed (CHISHOLM, 1957, pp. 177 ff.) that the familiar Pavlovian conditioning of a hungry dog, where after the conditioning the bell has come to be a *sign* of food, is one of the simplest kinds of psychological phenomena to require intentional concepts for an adequate explanation. Suppes has proposed a behavioristic definition for the notion of CS (conditioned stimulus, in the example the sounding of the bell), being a *sign* for US (unconditioned stimulus, the presence of food). *R* stands for the response (salivation). The dog, as you will remember, learns that the sounding of the bell is a sign of food in a process of trials, where in the beginning the dog salivates only in the presence of the food, but not upon the presentation of the bell. There is then

a series of training trials in which the food and the stimulus bell are presented simultaneously and the dog salivates. Finally, after these training trials, the dog responds by salivating upon presentation of the bell alone. Suppes' definition consists of the following four conditions:

- (1) $P(R_1|US_1) \approx 1$;
- (2) $P(R_1|(CS_1 \& \neg US_1)) \approx 0$;
- (3) For $1 < m \leq n$, $P(CS_m \& US_m) \approx 1$ & $P(R_m|(CS_m \& US_m)) \approx 1$;
- (4) For $n' > n$, $P(R_{n'}|(CS_{n'} \& \neg US_{n'})) \approx 1$.

Chisholm has formulated a number of difficulties that he thinks bars a behavioristic definition of "sign". Suppes' definition overcomes all of these. However, it is easy to give examples that satisfy Suppes' four conditions, but where we would *not* say that CS is a *sign* of US. Consider, for example, the case where we have two kinds of food for the dog, both in the form of small pellets. One kind of pellet has a smell and a color, say red, that attracts the dog. From the very beginning, the dog salivates when presented with these pellets. The other pellets have no attractive smell nor color, let us say they are green, and the dog has never shown any interest in them, even when hungry. Using an intentionalist vocabulary, we might say that the dog has never discovered that these pellets are food.

During a series of training trials the dog is now given a mixture of the two kinds of pellets. The dog is unable to separate the red pellets from the green ones and comes to eat them both. He thereby discovers that the green pellets, in spite of their unattractive appearance, taste very well, perhaps even better than the red ones, and after the training trials, he starts salivating upon presentation of the green pellets alone.

According to Suppes' definition, the green pellets have now become a sign for the red ones. However, this clearly conflicts with what we normally mean by the notion of a sign. Hence Suppes' definition is inadequate. The definition can easily be patched up, but again, new counter-examples may probably be found for the patched-up definition.

However, this brief discussion of Suppes was a digression. Let us return to our discussion of the more general problems that intentionality raise for behaviorism.

5. Attention and response

A second main place where intentionality comes in in connection with stimulus and response, is when we consider that the subject's registration of the stimulus and the responding are both actions. That they are actions

is brought out clearly by the following kinds of difficulties that are often discussed in psychological reports: In connection with the subject's registration of the stimulus, there is the problem of whether the subject attends to the stimulus or simply is staring at it without seeing—as William James put it in 1890: "staring at it in a vacuous, trance-like way"³.

If the subject does not attend to what is in front of him, there is of course no limit to what he might be thinking of and what caprice might prompt his response. And even if he does attend, what does he attend to? The person in front of him, his eyes, their color, their shape, or some object associated in some way or other, e.g. by memory, with one of these objects?

It is typical of the intentionalist view of action that an action is underdetermined by the external physical features of what goes on, e.g. the movement of the body and the direction of the eyes.

Then, next we have the problem that the response, too, is an action. Three questions immediately arise in connection with the response. First, what brings it forth? The stimulus, whatever the subject takes *that* to be, the subject's desire to impress the experimenter or to mislead him, or sheer caprice? And secondly, there is the experimenter's problem when he tries to describe the response. Even in as simple a case as the assent to or dissent from a sentence, Quine has acknowledged, in his reply to Hintikka in the volume *Words and Objections* (DAVIDSON and HINTIKKA, 1969, p. 312), that the decision whether to treat a piece of native behavior as assent or as something else, even perhaps as dissent, is a question on a par with the general problems of meaning and translation.

Thirdly, if the experiment involves verbal instructions, there is the problem whether the subject has understood the instructions, i.e. in which way he has interpreted them. Given, then, that intentionality seems to creep into even the simplest and most basic notion of behaviorism, it seems to me that one should take this into account when one wants to study man on the basis of his behavior.

6. Empirical evidence

The original aim of behaviorism was to base the study of man on what we can observe. That is, the study of man should be based on empirical evidence, evidence that reaches us through our senses. This aim is what

³ JAMES (1890), I, p. 222. Here quoted from NATSOULAS (1977), p. 80.

almost everybody would find appealing about behaviorism. There are many of us who would sympathize with Watson and others in their reaction against the excesses of mentalism and introspectionism in the beginning of our century. However, when we base our study of man on the evidence that reaches us through our senses, we must heed carefully what this evidence is. As we have noted, what we experience through our senses is not sensory data, far less stimulation of sensory surfaces. We experience physical objects, but not only that. Once we accede to Husserl that the impingements on our sensory surfaces under-determine what we experience, it seems quite arbitrary to say that the only objects of sensory experience are physical objects. We experience shapes and colors, and when it comes to man, we experience not just bodies and bodily movements. We may do that, but that happens only rarely. In most of our normal life with others, we experience persons and their actions, not bodies and their movements. This is not just a playful extension of the notion of experience. My point is that our basic experience of others, that which gives us evidence against which our psychological theories have to be tested, is experience of persons, not bodies, and of actions, not movements.

Similar points have been made before, by phenomenologists, Wittgenstein and others. Now, however, comes a most important point of difference between my view and that of many of the phenomenologists and Wittgensteinians. Although the evidence against which we have to test our theories is this kind of meaning-imbued experience, this does not mean that this experience is a rock-bottom foundation for science. There is no rock-bottom. Even though the impingements on my sensory surfaces may remain the same, what I experience may come to differ, as my beliefs and theories concerning the world change. What I now take to be veridical, I may tomorrow come to regard as misperception.

When it comes to psychology and the study of man, the situation is similar, but with even greater leeway for vacillations. I experience the other as a person, with beliefs and desires that are reflected in his actions. But I may be wrong in what I take his actions to be; even though his actions are what I perceive, there are many different possibilities for what his actions are, all of them compatible with the irritations of my sensory surfaces, and correspondingly for his beliefs and desires. Hence again, not only are my theories tentative, but so is also the evidence upon which they are based.

7. Behaviorism as an attempt to minimize the effects of the intentional

Given all this, I think that we may look upon many of the methodological precepts and techniques developed by behaviorists in a different way. Rather than considering various experimental set-ups and the like as means of eliminating the intentional, which, I have argued, is impossible, I regard them as ways of reducing the under-determinateness that invariably comes in due to the intentionality, in stimuli, in response, etc. The spectrum of different objects that we can possibly take a person to perceive, varies clearly with the various patterns of irradiation on the eye, etc. We noticed for example that the duck-rabbit picture was particularly ambiguous. Similarly, the spectrum of different actions that we can take a person to perform varies with the kind and complexity of the bodily movement that is taking place.

As our psychological theories improve, we may learn more and more about this and come to design better and better experiments. Already, in our present design of experiments, we are being helped by the intuitions and theories we presently have concerning man, his experiences and actions. This, by the way, is a reason why behaviorists easily come to delude themselves if they try to describe stimulus and response in a purely physical way. In the case of verbal responses, for example, one might try to give purely phonetic descriptions of what the subject utters instead of trying to interpret it in view of our whole tentative theory of the subject and his activity. Consider then in how many ways one can formulate basically equivalent responses in a language. Assent, for example, can be expressed by "yes", "right", "I agree", "you have learned your lesson well", "you amaze me", and so on, almost *ad infinitum*. The interconnection between all of these responses, which may make us want to treat them on a par if assent is all that we are interested in, would escape us completely if we just were to regard them as so many different phonetic sequences. Instead, of course, we regard the responses as meaningful expressions and group them in equivalence classes in view of our tentative theory of what the subject means and does.

8. What kind of theories are needed for the study of man

Now, finally, we come to the problem of how we go about finding out what people mean and do. Here, as elsewhere, we must start out from a tentative theory, which we should gradually try to improve and may also come to reject as our research goes on.

In studying man, we need a very comprehensive theory, which includes a theory of action, a theory of meaning and communication, a theory of reference, and a theory of knowledge. All of these are interconnected, in such a way that evidence is transmitted between the theories. The exact way in which this happens, we shall see as we go on.

9. Theory of action

Let us begin with the theory of action. Like many writers on the subject, for example Fodor in his recent book *The Language of Thought* (FODOR, 1975), I hold that action should be explained by the models of formal decision theory. That is, the agent conceives of himself as being in a situation where he has a set of options, or alternatives. He has beliefs concerning how his choice between the options will affect the likelihood of various consequences, and he attaches values, positive or negative, to these consequences. He then chooses the option that gives the highest expected utility.

Also game theory has to be brought in. The agent conceives of himself as being surrounded by other agents who in their turn conceive of him as an agent. His recognition of this and of the influence that his action will have on their actions affects his estimation of the probabilities of the various consequences that will ensue from his action.

However, when in this way I appeal to decision theory and game theory for an explanation of human action, one qualification is necessary. For the explanation of action, standard normative decision theory and game theory will not do. We are not all of us rational decision makers. What we need, is empirical decision theory and game theory, a theory of how people, as a matter of fact, make decisions. The theory we need for explanation of action will relate to normative decision theory the way a consistent psychologistic approach to reasoning would relate to logic. A psychologistic philosopher, if he is consistent, should find out how people in fact reason, what arguments they regard as valid and what kind of conclusions they tend to draw, however fallaciously, from a given set of premises. In our empirical decision theory we must take into account how people normally consider only a very small number of the alternatives for action that are open to them in a given situation, how they reflect only about some of their consequences, often have odd beliefs about the probabilities of the various consequences, and also may have values and

preferences that deviate considerably from those we think they ought to have.

We have to incorporate in our explanatory theory what we know about factors that systematically mislead people in their estimates of probabilities and about factors that influence their preferences and often make them fluctuate considerably and rapidly. We have to include information about an agent's training and past performance, about panic and other factors that may influence his rationality, and so on.

The empirical studies of decision making and choice by Tversky, Suppes and others⁴ are highly relevant here.

Clearly, our explanation and prediction of a person's actions will come to depend very much on our information and theories concerning this particular person, his characteristics as a decision maker, for example whether he shuns risks or enjoys them, and his conscious and unconscious beliefs and attitudes at the moment where the action took place.

To these other sources of information we shall return shortly. Let me, however, first make two observations, one concerning the status of the normative theory of rationality and one concerning the structure of explanation of action.

The normative theory of rationality is normative in the following sense, which is important in our context: If an action conforms to the normative decision theory, then this explains it. If it does not, then the deviation from normative decision theory has to be explained in order that we shall understand the action. Such deviation is normally explained by bringing in causal factors of the various kinds that I have just mentioned.

My second observation concerns the structure of the explanation of action: we should note that there is never just *one* reason, i.e. one desire or belief that enters into the explanation. However, as in other areas of explanation, when an explanation is given, we usually pick out one or just a few factors, that we give as *the* reason(s) for the action. Which factors we pick depends upon their relative weight and the circumstances of the explanation, for example, what factors are not already known to the person for whom the explanation is intended.

Hence, there seems to me to be no incompatibility between there being a number of reasons that enter into explanation by reason and our still giving only one, or a small number of these as *the* reason(s) for the action. In fact, the situation here seems parallel to the situation in causal ex-

See e.g. TVERSKY (1972), TVERSKY and KAHNEMAN (1974).

planation. In the case of causal explanation, too, one often just mentions one, or a small number of the causes of an event, e.g. those of which the person for whom the explanation is intended is not yet cognizant.

In fact, I think that the parallelism between explanation by reason and explanation by causes extends far beyond this. I will even suggest that rather than assimilating reasons to causes in the simple way that the instantiation of causal laws suggests, one should be aware that causal explanation, like explanation by reason, makes use of a whole intricate theory and not a single simple causal law. To take an example, let us consider the case where we explain why an iron ball is falling towards the center of the earth by saying that the ball is heavy and that we have a law of physics to the effect that every heavy body here on earth will fall towards the center of the earth. Clearly, this is sometimes true, sometimes false. If some other factor enters the picture, like a support that keeps the object in place, an electromagnetic field or the like, the ball will perhaps not fall, perhaps will it even move upwards.

We could, of course, save our law in the face of counterexamples like this by adding clauses to it, we might say: a heavy object here on earth which is not supported and not in a magnetic field will fall towards the center of the earth. Similarly in the case of actions, we might in the case of a person who opens the window begin with the following simple "law": "A person who wants fresh air and believes that he will get it by opening the window, will open the window". Noting the many exceptions from this "law", we might improve it to: "A person who wants fresh air and believes that he will get it by opening the window and who also does not believe that anybody in the room will suffer from the draft, will open the window". However, both in the example from physics and the example from action theory, our new revised "law" has exceptions. They both have to be supplemented with ever new clauses.

In action theory, we quickly see that there is little point in trying to formulate ever more complicated laws. Instead, we attribute to each agent a number of propensities towards action, a set of desires and beliefs, and we devise a decision theory which tells us how in a given situation all these propensities are fused into one resultant propensity for action. Similarly in physics we attribute to each object a number of propensities, mass, electric charge, a certain position relative to other masses and electric charges, etc., and on the basis of this we determine the resultant propensity for e.g. movement in the given situation. Another way of looking at this is that physical objects and events, like actions, can be

described in numerous ways, for each of which they instantiate different "laws", i.e. show different propensities. These laws, or propensities, by the way, need not be deterministic. They may be probabilistic, in action theory as well as in physics.

Formally, there are hence many similarities between the situation in physics and that in action theory. However, there are also important differences. One such is that in physics we have found a way of reducing the basic propensities to very few, mass and electric charge, position in gravitational and electromagnetic fields, etc., and one hopes for further reductions.

In action theory, the situation is much more complicated. We have to deal with values, like money, beauty, unspoiled nature, peace, love, etc., that we have not learned how to compare and that are perhaps incomparable.

We have also noted how these factors and the manner of their interplay may vary from person to person. Another factor, which also leads to differences between physics and action theory, is the capacity that the objects we study in action theory, viz. human beings, have to consider various possibilities. Some of these possibilities better satisfy our preferences than those we would reach if we were only to follow a gradient from where we are to some local maximum.

There are also further differences between action theory and physics, for example connected with our capacity for reflection, etc. There will not be time to discuss this here, but I have included some remarks on it in a paper in German that was published some months ago. (FØLLESDAL, 1979.)

10. Evidence for attributions of beliefs and values

The theory of action enables us to explain a person's actions once we know his beliefs and values. However, our problem is the opposite, we observe his actions, or rather what we take his actions to be, and we then try to form hypotheses concerning his beliefs and values.

This is the problem of so-called *revealed preference* which has been much discussed in economics. One particular difficulty here is that an action depends on both our beliefs and values in such a way that our determination of one of these factors depends on and changes with our assumptions concerning the other.

It has long been popular among economists to hold that the only way of determining a person's preferences is by examining his actual choices. However, there are more sources of information. First, we may simply ask him. One reason why economists stay away from this is that one cannot always trust what a person says. He may be lying, to us and perhaps to himself. He may sincerely believe that the reasons for his choice are different from what a person with more self-insight would think.

11. Speech acts as a species of actions

Many economists and philosophers would say that the ultimate basis for deciding what a person desires and believes is not what he says, but what he does. I do not think that there is just one basis, except to the extent that we consider also a person's speech acts as a form of action. Our theory of a person must account also for his speech acts. Our hypotheses concerning his beliefs and values must explain why certain of the things he says should not be trusted. Even a person's saying something that is false may serve to confirm our theory of him.

12. Epistemology and rationality

The hypotheses we make use of in such cases are not just hypotheses concerning the person's truthfulness and motives. Especially in cases where a person sincerely seems to believe something that we take to be false, this may fit in with our hypotheses concerning what he has perceived and not perceived, together with our epistemological theory of how a person's beliefs are formed and changed. Likewise, in attributing attitudes and values to a person, we have to make use of a theory of attitude formation and attitude change, together with hypotheses concerning the experiences, influences and development that the person has been through.

This theory of attitude formation has to contain theses concerning for example the need for consistency of preferences at a time (e.g. to what extent and why the preference relation has to be transitive), the consistency of preferences over time (what kind of preference changes should be expected in a person and what kind not; if any kind of rapid and capricious preference change were permitted, we would simply have too much leeway in our determination of a person's preferences on the basis of his actions). We also need assumptions concerning a person's concern for his own

future, in order to determine the "discount rate" he uses when he estimates the present value of future consequences of his actions.⁵

In view of this kind of considerations concerning the formation and change of a person's beliefs and attitudes, it should be clear that neither with regard to beliefs nor with regard to values should we simply "maximize" agreement. Rather, we should use our epistemology and all we know about a person's past experience to find out where we should expect and not expect agreement. This is clearly Quine's view when he emphasizes perception and agreement concerning trivialities and absurdities as a basis for translation. (QUINE, 1960, pp. 59–69.) It is also Davidson's view, as is particularly clear in his later writings.⁶ Although the maxim of maximizing agreement is important in Wittgenstein, in Gadamer and in many others, it should be relegated to a subordinate position; in a considerably weakened form the maxim is simply a consequence of the central importance of an epistemology when we study and theorize about persons as we do in psychology.

There is much to be added. I shall only now at the end mention two points, one having to do with ostension, the other with the so-called *introspective reports*.

13. Ostension

First: ostension. In view of our explanation of a person's actions we come to regard some situations where he is pointing as cases of ostension, that is, as attempts to indicate a reference. In such cases, the pointing gives us valuable information concerning what the person intends to refer to. Clearly, pointing does not uniquely determine a reference. Any physical object whose surface includes the first opaque surface in the direction of pointing is a candidate, so is any object related to one of these by the so-called *deferred ostension* (QUINE, 1969, pp. 39–41). However, we should definitely take the person as referring to one of these objects and not to something else. Ostension hence eliminates certain misinterpretations, and it should override our epistemological considerations concerning a person's beliefs. The epistemological considerations do after all

⁵ For a more thorough discussion of these and other factors, see ELSTER (1979).

⁶ DAVIDSON (1967, esp. p. 313; 1970, esp. p. 186; 1973, esp. p. 324 and note 14). Davidson makes his position with respect to agreement particularly clear in DAVIDSON (1973, esp. p. 19; and 1975, esp. pp. 20–22).

not single out just one interpretation as the only possible one. As in the case of scientific theories we should normally prefer the simplest of the alternatives. However, when this choice of interpretation comes into conflict with the evidence provided by ostension, we should give up the interpretation and side for one of the interpretations that are compatible with what we derive from ostension. The conflict may thereby in some cases even come to tell us that some of our epistemological assumptions are wrong.

14. Introspective reports

As for my final little point, concerning introspective reports, it will have been clear already from what I have been saying that I think that we should admit such reports. I have mentioned that we may ask a person about his preferences and his beliefs, and also how he feels, etc., and we then use his answers as data against which we test our theory. These data are of course not incorrigible, just as little as other data. The main point, which makes them relevant and useful, is that they do relate to our theories. They fit in and confirm the theory, or they do not fit in, and disconfirm it.

Note that I do not claim that all data are equally good, some are more corrigible than others. We have noted that observation of a person's choices often make us doubt his verbal reports on his preferences. Likewise a person's reports on his beliefs or on what something looks like to him may seem dubious in view of what he actually *does* do. However, as we have noted, observation of a person's actions is not a rock-bottom, either, one and the same bodily movement is compatible with the agent's performing any one of a number of actions. Listening to what he says and, of course, studying all the rest of his behavior may make us change our view on what action he performs.

Behaviorists have often been reluctant to admit introspective reports. In the case of some ontological behaviorists this may be due to a dogma that there is nothing there to inspect and report on. This dogma I have never seen supported by good arguments. Other, more epistemologically motivated behaviorists bar introspective reports for one of three reasons: the phenomena reported on cannot be intersubjectively observed, they involve intentional notions, or they are unreliable.

As for the phenomena reported on not being *intersubjectively observable*, this is admittedly so. However, as Alston has observed in a carefully argued article on private data (ALSTON, 1972), what matters in science

is not really intersubjective observability, but the possibility of independent testimony. And as I have argued all along in this lecture, there is a lot of such independent testimony available, the introspective reports and the observations of action have to fit in with one another, etc.

As for the second point, their *intentionality*, introspective reports do admittedly often concern intentional phenomena. However, as I have argued, intentionality seems to be unavoidable, even for the most ardent behaviorists, and once one has to depend on intentional notions, then it is methodologically wise to admit all the evidence that is relevant to these notions, also that yielded by introspective reports.

The third point, finally, concerning the *unreliability* of introspective reports, is important. Many introspective reports are highly unreliable, but others are not, and similarly differ also reports concerning external behavior. It is important for methodology that we have theories concerning the reliability of various kinds of data. Such a theory is, in the case of psychology, just a corollary of the general theory of man that psychologists seek to arrive at.

15. Conclusion

Behaviorists often proclaim that they are getting on well without intentionality or even that there is no such thing as intentionality. However, I have argued that intentionality permeates the objects and processes studied in psychology as well as the observations and procedures used in such studies, including those used by behaviorists. Further, I have argued that there is an interplay between the different fields of study: action, perception, etc., such that our explanatory hypotheses in one of these fields must fit in with those in the other fields. We have, for example, observed how the beliefs and values that we attribute to a person in order to explain his actions, must fit in with those we ascribe to him in order to interpret what he says. Evidence that bears on one of these fields thereby comes to bear also on the others. A theory of intentionality is in part a theory of the evidential interplay between these different fields. For behaviorism, with its emphasis on evidence, this interplay should be a major concern. From an intentionalist point of view, behaviorism may be looked upon as an endeavour to arrive at methods and experimental set-ups that give as reliable data as possible, especially by minimizing the ambiguities due to intentionality. By openly bringing in a theory of intentionality as a guide in this endeavour, we will get a better behaviorism.

References

- ALSTON, W. P., 1972, *Can psychology do without private data*, Behaviorism, vol. 1, pp. 71–102
- CHISHOLM, R., 1957, *Perceiving: A philosophical study* (Cornell University Press, Ithaca, N. Y.)
- DAVIDSON, D., 1967, *Truth and meaning*, Synthese, vol. 17, pp. 304–323
- DAVIDSON, D., 1970, *Semantics for natural languages*, in: *Linguaggi nella società e nelle tecniche*, pp. 177–188 (Milan)
- DAVIDSON, D., 1973a, *Radical interpretation*, Dialectica, vol. 27, pp. 313–328
- DAVIDSON, D., 1973b, *On the very idea of a conceptual scheme* (Presidential Address, American Philosophical Association, Eastern Division Meeting, Atlanta 1973), in: Proceedings at the American Philosophical Association 1974, pp. 5–20
- DAVIDSON, D., 1975, *Thought and talk*, in: *Mind and Language*, ed. Samuel Guttenplan, pp. 7–23 (Clarendon Press, Oxford)
- DAVIDSON, D., and J. HINTIKKA (eds.), 1969, *Words and objections; Essays on the work of W. V. Quine* (Reidel, Dordrecht)
- ELSTER, J., 1979, *Ulysses and the Sirens: Studies in rationality and irrationality* (Cambridge University Press, Cambridge)
- FODOR, J. A., 1975, *The language of thought* (Harvester Press, Hassocks, Sussex)
- FÖLLESDAL, D., 1979, *Handlungen, ihre Gründe und Ursachen*, in: *Handlungstheorien — interdisziplinär*, ed. Hans Lenk, vol. 2, pp. 431–444 (Fink, Munich)
- JAMES, W., 1890, *The principles of psychology* (Holt, New York)
- MARTIN, M., 1978, *Interpreting Skinner*, Behaviorism, vol. 6, pp. 129–138
- NATSOULAS, T., 1977, *On perceptual aboutness*, Behaviorism, vol. 5, pp. 75–97
- NAESS, A., 1936, *Erkenntnis und wissenschaftliches Verhalten*, Skrifter utgitt av Det Norske Videnskaps-Akademii i Oslo II. Hist.-Filos. Klasse, 1936, No. 1 (Dybvard, Oslo)
- QUINE, W. V., 1960, *Word and object* (The Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass.)
- QUINE, W. V., 1969, *Ontological relativity and other essays* (Columbia University Press, New York)
- QUINE, W. V., 1974, *The roots of reference* (Open Court; La Salle, Ill.)
- SUPPES, P., 1969, *Behaviorism*, in: *Studies in the methodology and foundations of science: Selected papers from 1951 to 1969*, pp. 294–311 (Reidel, Dordrecht)
- SUPPES, P., 1975, *From behaviorism to neo-behaviorism*, Theory and Decision, vol. 6, pp. 269–285
- TVERSKY, A., 1972, *Elimination by aspects: A theory of choice*, Psychological Review, vol. 79, pp. 281–300
- TVERSKY, A., and D. KAHNEMAN, 1974, *Judgement under uncertainty: Heuristics and biases*, Science, vol. 183, pp. 1124–1131
- VON UEXKÜLL, J., 1921, *Umwelt und Innenwelt der Tiere* (Springer, Berlin)
- VON UEXKÜLL, J., 1928, *Theoretische Biologie* (Springer, Berlin)

NEW PERSPECTIVES ON COGNITIVE PSYCHOLOGY

B. M. VELIKHOVSKY and V. P. ZINCHENKO

Moscow State University, Moscow, U.S.S.R.

Until quite recently American psychology was regarded as behaviouristic by most scholars in the world. Today, however, this notion accords with neither the self-awareness of the leading American psychologists nor the actual state of affairs established in experimental psychology during the last two decades. In the opinion of some specialists in logic, methodology and philosophy of science the latest development of American psychology—sometimes defined as a revolution (BOWER, 1975)—is entirely connected with the rise of a new psychological school—cognitive psychology (GROEBEN and SCHEELE, 1977; SEGAL and LACHMANN, 1972; among many others). Superficial evidence of the paradigm change is the rapid increase in the number of studies in the fields of perception, memory, thinking, attention and communication.

The new conception of subject and method in psychology is a more serious evidence of the change. The representatives of cognitive psychology, together with several linguists, argued against the mindlessness of the behaviouristic approach and tried to give a new answer to the old question of human behaviour determination. This question which is, as U. NEISSER (1976) says, “too important to be left to the behaviourists and psychoanalysts”, receives in cognitive psychology the following answer: human behaviour is determined by the knowledge which a person possesses. Thus, on this interpretation, cognitive psychology turns out to be quite closely related to Gestaltpsychology and earlier trends of mentalistic psychology. Indeed the term “neomentalism” introduced into the psychological vocabulary by A. PAIVIO (1975), could with justification be applied in the case also.

This interest in the organization of the internal representation of knowledge certainly has a strong intellectual appeal. In fact, to fulfil

general public expectations every school of psychology would more or less deal with the inner life of Man. However, this is elusive and fluctuating. Only great poets, writers and artists were able to describe this unstable and inconstant matter. The phenomenologists and existentialists also tried to stop, catch and describe the inner life, but most often their attempts were not successful and possessed little scientific reliability. Although Gestaltpsychologists brought about an important advance they, as a rule, failed to apply the experimental method *per se*. Therefore, what is of interest to us is their observations not their explanations. And explanation was what cognitive psychology attempted to provide.

As their main method of research cognitivists choose the method of operationalism which was also used earlier by neobehaviourism. Relying on modern technical means of laboratory experimentation cognitive psychology accumulated impressive experience in the hypothetico-deductive reconstruction of cognitive and lately also of executive processes.¹ This allowed one to give an operational content to the hypothetical intervening variables the very existence of which was only assumed by the neobehaviourists. By far the greatest number of interesting facts were obtained in such areas as psycholinguistics, the study of short-term mental processes and cognitive psychophysics, that is in the analysis of the decision-making processes. Nobody will deny now that a deeper understanding of the most important psychological phenomena is not possible without experimental investigation of their functional structure or microstructure.

Today cognitive psychology not only dominates the whole province of general psychology but is well represented in the field of differential and applied psychology as well.² In the English-speaking countries a purely geographical spread should also be noted (see, e.g., UECKERT and RHENIUS, 1979).

But whilst admitting an important contribution by cognitive psychology to the methodology and phenography of psychological experimentation

¹ Among the most important sources we should mention the three volumes edited by R. L. SOLSO (J. Wiley and Sons, as well as L. Erlbaum Ass., 1973-75), six volumes edited by W. K. ESTES (L. Erlbaum Ass., 1975-78) and the multivolume series *Attention and Performance* (North-Holland, Academic Press and L. Erlbaum As., 1967-78).

² See, for example, the book by EYSENCK (1977) as a demonstration of the cognitivistic approach to the problems of individual differences. The cognitive theories of personality and the variants of the so-called *cognitive psychotherapy* are as a rule less based on the methodological achievements of this school (KOZLOVA, 1976; MAHONEY, 1974). Many possibilities for the practical application of the methodology of cognitive psychology are reviewed in the multivolume series edited by V. ZINCHENKO (1970-79).

we wish to point out the difficulties which arise in the initial and most important area of cognitivistic research, that is psychology of perception and memory. In common with several other authors (e.g. NEISER, 1976, NEWELL, 1973) we think that the difficulties are of a fundamental nature and cannot be solved just by the accumulation of empirical data. The achievements of cognitive psychology are not sufficient to overcome all of the shortcoming of behaviourism and earlier schools of psychological thought.

If one tries to analyze the theoretical ideas developed in cognitive psychology, some remarkable contradictions will be apparent.³ The main paradox is that the representatives of this, on the whole, structural approach did not solve the problem of selecting adequate units of analysis. Usually one considers as such units different blocks of information-processing and operations which allow the transmission of information from one block to another. Usually the blocks are organized in sequential chains or more rarely in hierarchies. Every block is characterized by several parameters: the place in the whole structure, the informational capacity, the direction of the storage and the form of representation. Many models contain three kinds of blocks: the sensory registers, the short- and the long-term memories. One of the most popular combinations of these features is shown in Table 1.

Although similar blocks are still reproduced in articles, books and especially textbooks, it is perfectly clear to the many research-workers

Table 1. A structure of the human information processing process (after ATKINSON and SHIFFIN, 1968, LINDSEY and NORMAN, 1972)

| PARAMETERS | BLOCKS | | |
|------------------------------|-----------------------|----------------------|-----------------------|
| | ICONIC MEMORY | SHORT-TERM MEMORY | LONG-TERM MEMORY |
| Place in the whole structure | 1st | 2nd | 3rd |
| informational capacity | $\rightarrow\infty^*$ | 7 ± 2 | $\rightarrow\infty^*$ |
| duration of the storage | 0.3 s | 10/. 40 s | $\rightarrow\infty^*$ |
| form of representation | sensory code | verbal code | semantic code |

* This is a psychological infinity ($\gg 7 \pm 2$), not a mathematical one.

³ The last unfinished article by A. R. LUSIA (1978) was devoted to the analysis of the difficulties arising in the cognitivistic studies of memory.

in the field that they cannot be adequate "building-blocks" of human memory (e.g. NORMAN, 1978). Let us consider the most obvious destination of the sensory storage mechanisms (visual sensory register, very short-term visual memory or simply iconic memory) and the mechanisms of semantic storage (long-term memory or abstract amodal memory). On the one hand, the analysis of perceptual learning the performance of some kinds of psychophysical tests and the recognition capacities for elements of large sets of complex visual scenes show that sensory components of memory can be quite "long-term". In this respect they can successfully compete with verbal information contrary to the well-known Whorf-Sapir hypothesis. For instance, in our laboratory, K.-D. Schmidt and one of the authors of this article showed some years ago that ordinary subjects could quite successfully recognize the elements of a large set of semantically, but not perceptually, similar slides even several weeks after the initial exposure.⁴ The performance of the subjects was a function of the presentation duration, not the "on-on" interval. Thus, we can speak about long-term sensory memory (see also, PAIVIO, '78). On the other hand, if we turn to the very short-term psychological processes we should admit the important contribution of semantics at the early stage of perception. It was already shown in the 60's that semantic categorization can precede any kind of explicit or implicit verbalization (ZINCHENKO and VUCHETICH, 1970). Moreover, in the case of skillful reading the rough semantic categorization could precede not only phonological descriptions of a printed word, but even its distinct visual perception (see ALLPORT, 1977; SCHEERER, 1978).

The same difficulties arise if one tries to analyze the relations between blocks of short-term and long-term memory⁵ or to apply the boxological ideology to the explanation of some perceptual phenomena. For example, such a simple visual phenomenon as marking could be explained by the assumption that it is necessary for a stable perception that information reaches a certain higher block of a visual system, say, a block of "conscious representation". Here the marking is described as being like the "race" of two "stimulus horses". We would expect that the subject would perceive

⁴ In this study about 1.000 slides with views of modern standard city-blocks were used. They could not be differentiated on the basis of their verbal descriptions (VELICHKOVSKY and SCHMIDT, 1977).

⁵ One of the authors (WICKELGREEN, 1975) cited about 20 arguments for this distinction admitting at the same time their insufficiency.

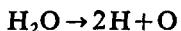
the stimulus which first reaches the box of conscious representation. However, from this point of view it is difficult to understand why the marking is backwards rather than forward, in other words why the second horse nearly always wins "the race".

If one goes beyond a few well-specified laboratory situations the task of analysis become extremely difficult. Almost all authors emphasize the hypothetical character of the block models of the more complex mental processes and their insufficient experimental verification. Some of them try to reduce the blocks to known neurophysiological structures, but others like WEISSTEIN (1973), write with irony about so-called *gnostic neurons*: the process of recognition begins with spot or line detectors and proceeds till the activation of the "my grandmother" or the "yellow Volkswagen" cell. In order to bring life into hypothetical block structures, D. NORMAN (1978), puts demons into them who make decisions and manage the circulation of the information. In this context the serious discussion of the place and role of a homunculus in the human cognitive process is not surprising (ATTNEAVE, 1960; POSNER, 1978). Although these works are among the most interesting one question arises: what is the value of the theories which use a homunculus to overcome their problems if one has to assume the existence of a second homunculus in order to explain the cognitive abilities of the first one and so on, ad infinitum?

Many representatives of cognitive psychology apparently believe that one can avoid these difficulties by introducing new blocks into the structural models or by the rearrangement of the old ones. Due to the operational, but not theoretical definition of most of the concepts in cognitive psychology it is a rather dangerous approach particularly when it became possible to carry out dozens of different experiments with the help of the same computer controlled display. It is hardly possible now to define *differentia specifica* of the following variants of at one time united mnema: active, working, operative, acoustic, verbal, articulatory, motor, iconic, echoic, semantic, episodic, etc. memory. In their turn those blocks dependent on author and experimental procedures are filled with concepts, attributes, markers, images, labels, nodes, names, surface and deep structures, programs of motor instructions, associations, propositions, rules, traces, etc. (cf. EYSENCK, 1977). But all these attempts "to improve" cognitivist theories will remain "New looks through old holes" if they do not touch the philosophical bases of this approach.

This base is in the overwhelming majority of cases of the well-known sensuous conception of mind. In particular, iconic memory with respect

to its characteristics and role in cognitive processes resembles suspiciously the retinal image of the associationistic psychology of the XVIII-XXth centuries. That is why one of the authors who made an especially valuable contribution to the development of the cognitive approach in the field of visual perception recently wrote: "Certainly Helmholtz, returning to the scene after an 80-years leave, would find himself on familiar terrain" (EPSTEIN, 1977, p. IX). But if the starting point of the theoretical analysis in psychology is the retinal image, a very difficult problem arises namely how to explain the obvious meaningfullness of our inner life? It is very naive to think that we can put consciousness into the boxes of cognitive psychology. One of the founders of Soviet psychology, L. VYGOTSKY (1934) wrote that theoretical analysis in psychology must not destroy the internal relations of the object under investigation. The units of analysis must preserve the main properties inherent in the whole and must be capable of development. He illustrated this idea with as an example the analysis of water. If the analysis of water goes deep enough:



we shall obtain elements (not "units") which have absolutely different properties, for instance with respect to fire.

In the following we shall try to show that the genetically prior unit of psychological analysis is not sensation, reaction, block, etc., but object action (or in German = *gegenständliche Handlung*). The ontological schemes of marxist psychology were always built on the basis of the primacy of the subject's practical activity (in German: *Tätigkeit*). In one of his works, V. LECTORSKY (1977) has shown specifically that this principle allows for the reconstruction of the development of the various forms of human activity. Concrete theories of these processes were also developed by A. LEONTJEV (1959) and more recently by K. HOLZKAMP (1973).⁶ Directed to the solving of real life problems the subject's activity for the first time gives rise to the appearance of mental reflexion. Simultaneously the activity is orientated and controlled by the reflexion. In order to accomplish this function, the mental reflexion must be objectively true. What are the consequences of this position for the specific analysis of perceptual and mnemonic processes?

* On such specific approaches see also RUBENSTEIN (1946), DAVYDOV (1972), ZINCHENKO and GORDON (1976), and others.

First of all one must insist that the perceptual image is at all stages of its formation something that connects, not separates, the subject and the world of objects around him. In fact in Phylogenesis a monadecacy of the mental reflexion would simply lead to the biological elimination of the particular species. The recent experimental research on the ontogenesis and microgenesis of perception has consistently shown that the development of perception does not obey the classical scheme: "monadecate image (retinal image, icon, visual field, etc.) → adequate image". On the contrary, the mainstream development of perception is in both these cases the transition from the globally adequate reflexion to the reflexion which is adequate also in details. In particular, research on the microgenesis of visual perception has shown that it begins with fast localization of the quasi-object regions in three-dimensional space and time, then the assessment of their general outlines follows and finally the invariant description of the internal geometry of an object takes place. These processes take about one third of a second, that is a time which is equal to the duration of one visual fixation (VELIKHOVSKY, 1977). Thus the percept corresponds to the three-dimensional world, not to the retinal image even if the image remains fixed as it is during one eye fixation. Still more clearly the irrelevance of the local sensory metrics is demonstrated in the studies of the holistic perceptual processes which take place in the macrointervals of time. These studies have been conducted by Soviet psychologists since the beginning of the 30's. The analysis of tactal perception has a special importance here, because in this case the very idea of some picture of sensory stimuli as the ground of perception loses any sense. (ANANJEV *et al.*, 1959.) As these questions are comprehensively discussed in the psychological literature it is not necessary to dwell upon the topic any further. Now we shall turn to arguments of another kind which are based on the works of some eminent physiologists.

It is well known what importance was attached to the subject's movements and actions by Sechenov. Another eminent physiologist, Sir Charles Sherrington, wrote: "We can suppose that the realization of actions directed to special goals opens up in the course of selection a possibility for elements of memory and anticipation to develop the psychological ability of "unfolding" the present back in the past and forward into the future. This ability is a necessary feature of the more advanced level of intellectual development of higher animals". (SHERRINGTON, 1969, p. 314.)

Similar views on movements and actions were experimentally developed later on by such Soviet physiologists and psychologists as N. VVEDENSKY,

A. UCHTOMSKY, N. BERNSTEIN, P. ANOCHIN and A. R. LUSIA. In their works, goal-directed movements were revealed as heterogeneous units which include elements of perception, memory, anticipation and control. They also include the emotional-evaluative attitudes of the subject. That means that goal-directed movements cannot be considered solely from the point of view of changes of spatial position. Their transformation into object actions is accomplished by the inclusion of an object in the movement as one of the components of the latter. Thus the movement becomes mediated by the object and its relations to other objects. For example, some tools of activity are progressively included in movements to such an extent that phenographically we experience an expansion of our tactual sensitivity beyond their normal borders to the ends of the tools (see, e.g., BERNSTEIN, 1966).

This characterization of movements differs completely from the interpretation of behaviourists who constrained themselves to the study of directly observed features. One gets the impression that cognitive psychology starting in opposition to behaviourism has moved to another extreme since it isolated cognitive processes from actions (see also WEIMER, 1977).

Certainly we would need to undertake a great deal of research in order to explain the transformation of different kinds of activity into relatively independent cognitive and personal structures. Our conception of the genesis of mental reflexion in interaction with the changing world is far from complete but its further development is covered for by the whole history of European and Soviet psychology. We mean first of all the interpretation of perception, memory and thinking as systems of perceptual mnemonic and intellectual operations; for example, in one of the first studies of memory from this point of view P. ZINCHENKO (1939) wrote that remembering as a special action assumes in its development different forms dependent on the components of this mental act: the object, the goal, the motive and the means (or operation) of remembering. A change in even one of these components leads to the transformation of the whole structure of the mnemonic action. Zinchenko emphasized that in the action the object appears not as "pure" stimulus abstracted from the intentions and knowledge of the subject, but as an essential moment of the action, as the subject enters into meaningful relations with the object.

According to this interpretation the informational content of memory cannot be completely dissociated from the meaning content. It should be stressed that we take into consideration not only linguistic meanings but also object meanings (in German: *gegenständliche Bedeutungen*). This

term was introduced into psychological literature by HOLZKAMP (1973) who emphasized that the experience of a subject's practical activity is more rich, especially on the earlier stages of development, than the system of verbal categories. In fact, some new experimental research shows that images can be organized into more or less stable systems of relations functioning along with the verbal categorical system in the process of decision of various practical and cognitive problems. It is interesting to note that the interpretation of memory representation as a multidimensional space (which is common to several contemporary approaches) sets forth the problem as being perfectly analogous to the problem of coordinated control of motor actions. Both in the recognition and in the control of movements the main problem is the restriction of surplus degrees of motion freedom in space appearing either in the form of entire environment, and the body scheme or in the form of the semantic space of denotative and connotative features (VELIKHOVSKY, KAPITZA and SCHMELJOV, in press).

The genetic and functional connections between mental representation and object action also clearly appear in the analysis of mental dislocations, rotations and transformations since they depend not only on perceived (imagined) cinematic properties of objects but also on their actual dynamical properties as well. This fact was discovered in our laboratory by B. BESPALEV (1976). It is obvious that the dynamical, not simply cinematic properties are of particular significance for the performance of spatial manipulations with real objects (see also RUNESON, 1977; ZINCHENKO and VERGILES, 1969).

Certainly all this poses the problem of unified description both for coordinated motor control as well as for coordinated cognitive processes. It seems to us that the old idea of polyphonic coordination proposed by Soviet linguist M. BACHTIN (1929) and the new hierarchical approach of TURVEY (1977), WEIMER (1977) and others could be rather promising in this respect.

Finally, we shall sum up our article. Despite an important contribution to the methodology and phenography of psychological experimentation cognitive psychology has failed to solve a number of theoretical problems the most important of which is the singling out of the genetically prior unit of analysis. On the whole, the antibehaviouristic tendency of cognitive psychology has found its expression in a return to the position of mentalistic psychology. The next perspective in the study of the cognitive as well as motor processes consists in the real overcoming of both be-

haviourism and mentalism. It is precisely in this direction that of late several trends in psychological thought inside and outside the cognitive approach have been developing.⁷ Such a revision is undoubtedly possible on the basis of the psychological theory of activity.

References

- ALLPORT, D. A., 1977, *On knowing the meaning of words we are unable to report: The effect of visual masking*, in: Attention and performance VI, ed. S. Dornić (L. Erlbaum Ass., Hillsdale, N. Y.)
- ANANJEV, B. G., L. M. VEKKER, B. F. LOMOV, and A. V. YAZMOLENKO, 1959, *Osjasanije v processah poznanija i truda* (Touching perception in cognition and at work, in Russian) (Pedagogika, Moscow)
- ATTNEAVE, F., 1960, *In defence of homunculi*, in: Sensory communication, ed. W. Rosenblith (MIT Press, Cambridge, Mass.)
- BACHTIN, M. M., 1929, *Problemi tvorichestva Dostoevskogo* (Problems in the study of Dostoevsky's creativity, in Russian) (Pisatiel, Moscow)
- BERNSTEIN, N. N., 1966, *Ocherki po physiologii aktivnosti i physiologii dvigenija* (Essays on the physiology of activity and physiology of movements, in Russian) (Medicina, Moscow)
- BOWER, G. H., 1975, *Cognitive psychology: An introduction*, in: Handbook of learning and cognitive processes, ed. W. K. Estes (L. Erlbaum Ass., Hillsdale, N. Y.)
- DAVIDOV, V. V., 1972, *Vidi obobshchenij v obuchjenii* (The kinds of generalization in learning, in Russian) (Pedagogika, Moscow)
- EPSTEIN, W., 1977, *Vorword*, in: Stability and constancy in visual perception, ed. W. Epstein (J. Wiley and Sons, New York)
- EYSENCK, M. W., 1977, *Human memory: Theory, research and individual differences* (Pergamon Press, Oxford)
- GROEBEN, N., und B. SCHEELE, 1977, *Argumente für eine Psychologie des reflexiven Subjekts* (Steinkopf, Darmstadt)
- HOLZKAMP, K., 1973, *Sinnliche Erkenntnis-Historischer Ursprung und gesellschaftliche Funktion der Wahrnehmung* (Fischer Athenäum, Frankfurt/Main)
- KLIX, F., 1971, *Information und Verhalten* (Akademik Verlag, Berlin)
- KOZOVA, J. N., 1976, *Lichnost kak sistema konstruktorov* (Personality as a system of constructs, Russian), in: *Sistemnyje issledovaniya* (System's researches), (Nauka, Moscow)

⁷ This evolution is best observed in the works of U. NEISSER. One need only to compare two of his books (1967; 1976) in order to see the importance he now attributes to the subject's exploratory activity. The followers of J. J. Gibson are especially active in this respect (e. g. MACE, 1977). There are also several researches comparing the foundations of cognitive psychology with the principles of marxist psychology which have been completed by such well-known psychologists as F. KLIX (1971), J. LINHART (1976), and M. STADLER (1979).

- LECTORSKY, V. A., 1976, *Principi predmetnoj dejatelnosti i marxistskaja teoria poznaniya*. V.: *Ergonomika*. Trudi VNIITE (Principles of object activity in the marxist theory of cognition. In: Ergonomics. Proceedings of VNIITE), Moscow; VNIITE, vol. 10
- LEONT'EV, A. N., 1959, *Problemi razvitiija psichiki* (Problems in the development of psychological processes, Russian) (MGU, Moscow)
- LINDSEY, P., and D. NORMAN, 1972, *Human information processing* (Academic Press, New York)
- LINHART, J., 1978, *Cinnost a poznavani* (Academia, Praha)
- LUSIA, A. R., 1978, *Paradoxi pamjati* (Paradoxes of memory, in Russian), Vestnik MGU. Psykhologiya, vol. 1, pp. 3-11
- MACE, W. M., James J. GIBSON, 1977, Strategy for perceiving: *Ask not what's inside your head, but what your head's inside of*, in: *Perceiving, acting, and knowing*, eds. R. Show and J. Bransford (L. Erlbaum Ass., Hillsdale, N. J.)
- MAHONEY, M. J., 1974, *Cognition and behavior modification* (Ballinger, Cambridge)
- NEISSER, U., 1976, *Cognition and reality* (W. H. Freeman, San Francisco).
- NEWELL, A., 1973, *You can't play 20 questions with nature and win*, in: *Visual information processing*, ed. W. G. Chase (Academic Press, New York)
- NORMAN, D., 1979, *Personal communication*, December
- PAIVIO, A., 1975, *Neomentalism*, Canadian Journal of Psychology, vol. 29 (4), pp. 536-541
- PAIVIO, A., 1978, *A dual coding approach to perception and cognition*, in: *Modes of perceiving and processing information*, eds. H. L. Pick, Jr., and E. Saltzman (L. Erlbaum Ass., Hillsdale, N. Y.)
- POSNER, M. I., 1978, *Chronometric explorations of mind* (L. Erlbaum Ass., Hillsdale, N. Y.)
- REYNOLDS, A. G., and P. W. FLAGG, 1977, *Cognitive psychology* (Wintrop, Cambridge, Mass.)
- RUBENSTEIN, S. L., 1946, *Osnovi obshchej psikhologii* (Foundamentals of general psychology, Russian) (Pedagogika, Moscow)
- RUNESON, S., 1977, *On visual perception of dynamic events* (University of Uppsala, Uppsala)
- SCHERER, E., 1978, *Probleme und Ergebnisse der experimentellen Leseforschung*, Zeitschrift für Entwicklungspsychologie und Paedagogischen Psychologie, vol. 10, pp. 347-354
- SEGAL, E. M. and R. LACHMAN, 1972, *Complex behavior or higher mental process*, American psychologist, vol. 27 (1), pp. 46-55
- SHERRINGTON, Ch., 1969, *The integrative action of the nervous system* (Nauka, Leningrad)
- STADLER, M., P. SCHNAB und TH. WEHNER, 1979, *Kognition als Abbild und Plan des Handels*, in: *Komplexe menschlicher Informationsverarbeitung*, H. Ueckert und D. Rhenius (Hrsg.), Beiträge zur Tagung "Kognitive Psychologie" in Hamburg 1978, Bern, Stuttgart (Hans Huber, Wien)
- TURVEY, M. T., 1977, *Preliminaries to a theory of action with reference to vision*, in: *Perceiving, acting, and knowing*, eds. R. Show and J. Bransford (L. Erlbaum Ass., Hillsdale, N. Y.)
- UECKERT, H., und D. RHENIUS (Hrsg.), 1979, *Komplexe menschlicher Informationsverarbeitung*, Beiträge zur Tagung "Kognitive Psychologie" in Hamburg 1978, Bern, Stuttgart (Hans Huber, Wien)
- VELIKHOVSKY, B. M., 1977, *Zritel'naja pamijat i model: pererabotki informatii chelovekom* (Visual memory and some models of human information processing, Russian), Voprosy psichologii, vol. 12 (6), pp. 49-61

- VELICHKOVSKY, B. M., M. S. KAPITZA, and A. G. SCHMELJOV, *Memory structure: From blockdiagramms to multidimensional spatial models*, in: Cognition and memory, ed. F. Klix (J. Wiley and Sons, New York)
- VELICHKOVSKY, B. M., and K.-D. SCHMIDT, 1977, *Perceptivnaja dolgovremennaja pamjat* (Perceptual long-term memory, Russian), Vestnik MGU. Psichologija, vol. 1 (1), pp. 35-44
- VIGOTSKY, L. S., 1934, *Mishlenie i rech* (Thinking and language, Russian). (SOCEKGIZ, Moscow)
- WEIMER, W. N., 1977, *A conceptual frame work for cognitive psychology: Motor theory of the mind*, in: Perceiving, acting, and knowing: Toward an ecological psychology, eds. R. Shaw and J. Bransford (L. Erlbaum Ass., Hillsdale, N. Y.)
- WEISSTEIN, N., 1973, *Beyond the yellow Volkswagen and the grandmother cell: A general strategy for the exploration of operations in human pattern recognition*, in: Contemporary issues in cognitive psychology: The Loyola symposium, ed. R. L. Solso (J. Wiley and Sons, New York)
- WICKELGREEN, W., 1975, *The short and the long of memory*, in: Short-term memory, eds. D. Deutsch and J. Deutsch (Academic Press, New York)
- ZINCHENKO, P. I., 1939, *Neproizvolnoe zapominanie* (Unvoluntary remembering, Russian), Nauchniye zapiski instituta inostrannich jazikov (Harkov)
- ZINCHENKO, V. P., (ed.), 1970-79, *Ergonomika*, Trudi VNIITE (Ergonomics, Proceedings of VNIITE), vol. 1-17
- ZINCHENKO, V. P., and N. Yu. VERGILES, 1969, *Firmirovanije zritel'nogo obraza* (Formation of visual image, Russian) (MGU, Moscow)
- ZINCHENKO, V. P., and G. G. VUCHETICH, 1970, *Scanirovaniye posledovatel'no fixiruemih sledov v kratkovremennoy pamjati* (Scanning of the sequentially fixed traces in the short-term memory, Russian), Voprosy psichologii, vol. 1, pp. 17-31

THE EQUILIBRIUM CONCEPT IN ECONOMICS

E. MALINVAUD

I. N. S. E. E., Paris, France

Economic phenomena are notoriously complex and changing. Experimentation about them is hardly feasible. These unfavorable features do not prevent a scientific approach in economics. On the contrary, they impose to economists adherence to strict methodological principles and they require that the accumulation of economic knowledge use clear conceptual frameworks.

Economics is not an integrated science. Categories are distinguished within economic phenomena, each category being the object of a particular branch of economics, i.e. of a particular "theory". For instance, the theory of prices and resource allocation studies how are determined the relative prices, the productions and the consumptions of the various goods; the theory of employment tries to explain the degree of utilization of productive capacities and in particular the rate of unemployment of the labor force; and so on. At the present stage of the development of economics, the theories concerning the respective categories of phenomena are not well articulated with one another, in the same way as approximate images of the same object viewed from different angles are usually not well articulated with one another. Moreover, the division of economic phenomena into categories studied by various theories is changing through time as science progresses with a finer analysis of each phenomenon and a better understanding of articulations between distinct phenomena.

Most economic theories are and will be built around the study of "equilibria". The precise definition of equilibrium varies from one theory

to another, but the basic concept remains the same. This paper will attempt to describe what this concept is, how it is used in some important branches of economics and why it is expected to be appropriate for the representation of reality.

1. The definition of equilibrium

The word *equilibrium* seems to have been first introduced in economics within the theory of prices and resource allocation. L. WALRAS (1874) gave it a prominent place; when the supply of a good is equal to the demand for it, the corresponding market is said to be "in a stationary state or in *equilibrium*" (p. 85); when on all markets supplies equal demands, then a "general equilibrium" holds (p. 157). A. MARSHALL (1890) spoke of "temporary equilibrium" (p. 331) in order to insist on the fact that the conditions for equality between supplies and demands were changing through time.

But the word was later used outside of this particular context. J. M. KEYNES (1936) and still more his disciples studied how the "unemployment equilibrium" depended on monetary and fiscal policies. Since then the progress of mathematical economics helped to a rapid spreading of the same abstract concept to practically all branches of economics.

For a long time the word "equilibrium" was associated with the idea of an economic order that would best satisfy human needs. For instance, A. MARSHALL (1890) spoke of "equilibrium between desire and effort" (p. 331). This association explains why the word, but not the concept, was refused by critics of the prevailing economic institutions (K. Marx did not speak of equilibrium but presented a theory of value that is best expressed in modern terms as a theory of equilibrium). In this respect the relevance of the theory of the general competitive equilibrium was and still is at the center of ideological debates, which interfere with the wide recognition of the explicit or implicit role of the equilibrium concept in most types of economic analysis.

A careful examination of this concept was made by F. MACHLUPI (1958). Writing under the title *Equilibrium and disequilibrium*, he first noted that these words were used in many theoretical and empirical contexts, but insisted on the role of equilibrium as an abstract notion that is always intimately linked with the model in which it occurs ("The model as well

as its equilibria are, of course, mental constructions", p. 54). He proposed the following definition for equilibrium: "a constellation of selected interrelated variables so adjusted to one another that no inherent tendency to change prevail in the model which they constitute" (p. 54), or more simply "mutual compatibility of a selected set of interrelated variables of particular magnitudes" (p. 55).

These definitions stress mutual compatibility but do not refer to the subject matter of economics. One may be more specific and stipulate that an economic equilibrium always recognizes the existence of agents who produce, trade, consume, lend, subsidize... The compatibility has two dimensions that are always present, even if only implicitly in some models: (i) the various actions of an agent must be compatible with one another and compatible with the constraints imposed on him, as well as with the aims he tries to achieve; (ii) the actions of the various agents must be mutually compatible: a trade for instance is a purchase for an agent and a sale for another. Indeed, the interdependences between the various agents and the various operations are central to any economic analysis.

These considerations motivate the following definition. *In the abstract representation of some category of economic phenomena, an equilibrium is a state in which the actions of various agents are mutually consistent with one another and individually compatible with the behavior of these agents.*

It is noteworthy that this definition, as well as the second one proposed by F. Machlup, do not make reference to a process that would lead to the realization of equilibrium or would maintain it. Speaking of equilibrium is, however, accepting the idea that such a process should or may exist. Indeed, in some economic theories, but not in all of them, attempts at making the process explicit and at studying it do exist. The third section of this paper will consider where economists stand in the justification of their implicit claim at considering equilibria as motionless states. The second section will on the contrary consider various types of equilibria independently of the processes that could enforce them.

Whatever the definition, the operational use of the concept is to permit comparisons between equilibria resulting from different sets of conditions. If some change in the environment or in economic policy is contemplated, predictive statements concerning the effect of this change follow from the comparison between the new equilibrium, in which this change has been introduced, and the corresponding old equilibrium. Hence, the concept is truly central.

2. Equilibria in various theories

One may get a flavor of the methodology of economics by considering briefly how an equilibrium is actually defined in some simple versions of a few important theories. This presentation has of course to hide a number of difficulties and will not do justice to the mathematical sophistication now achieved in economics, a sophistication that was found necessary for an appropriate treatment of the subject matters.¹

(i) In order to explain how unemployment varies from one year to the next and how economic policy can act on it, one is usually satisfied with a macroeconomic analysis in which agents are not individually identified but collectively represented and in which goods are similarly aggregated. The argument runs about as follows.

Unemployment may be viewed as resulting from a discrepancy between the supply of labor by individuals and the demand for labor by firms. The latter follows directly from the level of production that is chosen and indirectly from the demand for goods that this level reflects. But the demand for goods depends on real incomes, which themselves appear for the most part as a result of production.

The theory that permits a precise analysis of these interdependences recognizes at least three groups of economic agents: firms that invest, employ and produce; individuals that work, receive incomes, consume and save; government that decides on the public demand for goods and on transfer payments. Whereas the problem usually is to compare various government decisions, the behaviors of firms and individuals must be represented together with the constraints that make them mutually compatible.

The simplest model for this theory assumes that the supply of labor by individuals is fixed and the demand for labor by firms is a given increasing function of their output Y , so that unemployment is inversely related to Y . Moreover, investment I is assumed to be predetermined and individuals real income to be equal to production Y . The consumption of goods by individuals is taken to be the function $f(Y)$ of their real income, this being a representation of their behavior. If G is the demand for goods by government, the mutual compatibility between supply and demand of

¹ To get a flavor of this sophistication, the reader may consider the recent survey by G. DEBREU (1979) on the existence of the competitive equilibrium.

goods in this model, which ignores foreign trade, is given by:

$$(1) \quad f(Y) + I + G = Y.$$

An equilibrium is a solution Y of this equation, in which I and G appear as independent variables. A contemplated policy change from G to G' is supposed to induce through (1) a change in output from Y to Y' , and correspondingly a change in unemployment. (This assumes that (1) has a unique solution in both cases.)

(ii) Technological progress, changes in tastes or in natural resources induce changes in the techniques of production, as well as in the levels of production and consumption of the various goods; simultaneously relative prices and wages shift. The theory of prices and resource allocation, which must explain this evolution and how it reacts to taxation policies or to regulation, must identify as many commodities h as there are distinct goods and types of labor ($h = 1, 2, \dots, l$).

The simplest version of this theory considers the case where each commodity has a price p_h , which is the same for all trades concerning this commodity and at which each individual agent believes that he could sell or buy as much as he would want: this is the core of the perfect competition hypothesis. Agents are usually divided into two groups: the consumers ($i = 1, 2, \dots, m$) and the producers ($j = 1, 2, \dots, n$). The actions of each one of them depend on prices, i.e. on the price vector p , and of course on the constraints and motivations that condition his behavior. The actions of all of them must be mutually compatible, because, for each commodity h , the total of consumption must be equal to the total of production plus whatever quantity ω_h , if any, was a priori available.

Let x_{ih} and y_{jh} , respectively, be the consumption of h by i and its production by j (these quantities are negative for a labor h provided by consumer i or for an input of h used by producer j). The theory explains how x_{ih} and y_{jh} depend on p . Here we may simply write the resulting functions $x_{ih}(p)$ and $y_{jh}(p)$. The mutual compatibility conditions then are:

$$(2) \quad \sum_{i=1}^m x_{ih}(p) = \sum_{j=1}^n y_{jh}(p) + \omega_h \quad \text{for } h = 1, 2, \dots, l.$$

An equilibrium is a vector p (with the corresponding x_{ih} and y_{jh}) that fulfills these l equations. A change in the exogenous resources ω_h is supposed to induce the change of p that system (2) implies.

(iii) The two preceding cases correspond to theories that have become

quite classical and whose properties have been extensively studied. The phenomena they represent are clearly interconnected since consumption and production explicitly appear in both cases. To connect the two theories is, however, not an easy task and the problem how to do it was the subject of many debates during the last forty years. It is fair to say that the connection is now rather well understood but that no general theory exists that would simultaneously explain changes in unemployment and in relative prices. This lack of a general theory is itself in part due to the difficulties of the theory of imperfect competition, difficulties to which we shall turn our attention in a moment.

The same abstract concept of equilibrium is also used for development of economic thinking in new directions. A noteworthy example appears with the present attempts to take full account of the formation of expectations.

It has been recognized for a long time that the expectations held by agents are important factors for an understanding of the economic situation and of its evolution. It is also more and more realized that expectations depend on what was observed about this evolution. A complete knowledge of many phenomena therefore requires a correct appreciation of the interdependence between the formation of expectations and other aspects of economic activities.

This interdependence is not easily mastered since it implies such difficult components as: the randomness of many phenomena, the fact that agents have very variable and always partial information, the frequent irrationality of human behavior in face of uncertainties. These difficulties culminate for the study of short or medium term evolution.

But for studying some aspects of long term growth, such as the reactions of people confronted with sustained inflation, or some problems of economic organization, such as those concerning financial markets, it is feasible and fruitful to define a "rational expectation equilibrium". In such an equilibrium agents form probabilistic (or even in some cases sure) expectations that are consistent with the information they receive; simultaneously they take actions that are consistent with these expectations; the actions of the various agents are assumed to be made mutually compatible, which determines the economic situation and the information about it that is made available to each agent. Depending on the case considered, such an equilibrium will have different features; it is sometimes taken as a "perfect foresight equilibrium" but it is always assumed to have, in a sense, "self-fulfilling expectations".

(iv) If the concept of equilibrium is pervasive in economic theory, it may also be quite challenging. It is particularly so for a very important aspect of economic phenomena: the fact that, in many cases, a small number of large firms or institutions each have a significant power on the situation that will obtain; each one of them, when deciding, takes account of the impact that its decision will have on the others and anticipates their reactions.

To deal with this aspect of reality, a frequent approach was to consider it as a deviation from perfect competition. But this approach cannot go very deep into the problem. It was rightly pointed out by J. VON NEUMANN and O. MORGENSTERN in their *Theory of Games and Economic Behavior* (1944) that a different logic is required, which would recognize from the beginning the feedback from the other players that any action of one of them would provoke, which would also recognize the possibility of coalitions between agents.

The main logical problem was then to formalize correctly the features of a situation in which the actions of the various players, if necessary grouped into coalitions, would be mutually consistent with one another as well as individually compatible with the aims and constraints of each player. The problem was then indeed to define precisely what an equilibrium had to be.

Surveying the various definitions proposed and used in the theory of games, or equivalently in the theory of imperfect competition, would lead us too far away from our subject. Suffice it to know that no single definition has been found fully appropriate. The von Neumann-Morgenstern solution, the core, the Shapley value, the nucleus, the nucleolus... are distinct concepts of equilibrium, usually leading to distinct solutions. This is a field in which scientific thinking has proved so far unable to propose a well defined and compelling answer to what is a well defined problem.

3. Equilibrium as a motionless state

The word *equilibrium* suggests a state that, if left alone, would not move. In economics an equilibrium is a state in which mutual consistency of the various actions has been achieved. One may accept, as an intuitive idea, the notion that such a mutual consistency is a precondition for the lack of movement.

But accepting the idea cannot be the end of the story. Indeed, if econ-

omists use equilibria as the central tool for their apprehension of real phenomena, they must also consider that these equilibria would spontaneously tend to materialize if unchanged conditions would prevail. It remains, however, to know in each case what process will enforce the equilibrium.

That such a question had to be faced was well understood since the concept of equilibrium has been used. Indeed, L. WALRAS (1874) insisted systematically on it. For instance, after discussing the equations for a general equilibrium in the exchange of several commodities, he writes: "Now let us see in what way this problem of the exchange of several commodities for one another to which we have just given a scientific solution is also the problem which is empirically solved in the market by the mechanism of competition". (p. 169.)

In other words, a complete theory requires not only a precise characterization of the relevant equilibrium but also a definition of the process that is supposed to lead to this equilibrium and a proof that the process does converge to it. In this respect, most branches of economics are rather weak. The process is often only hinted, in terms that are vague enough to permit many interpretations. When it is precisely defined, it is often presented as one possible candidate among others that could as well pretend to represent the phenomenon; it is then meant to be illustrative more than realistic. The proof of the convergence to equilibrium often involves special conditions, which are not even always formulated on the basic elements of the theory and are seldom empirically tested.

(i) For the unemployment macroeconomic equilibrium of equation (1), the textbooks suggest that the distribution of incomes to consumers and the formation of consumption demand require a lag of one period, so that consumption is $f(Y_{t-1})$ while production is Y_t . The process is therefore defined as:

$$(3) \quad Y_t = f(Y_{t-1}) + I + G.$$

Convergence is then easily proved from natural and well tested properties of the consumption function f , namely that it is larger than Y for very low values of Y and that it has a positive derivative that is bounded below 1.

This is of course a very simplified picture of reality, as the equilibrium of equation (1) was. The picture has, however, been made more realistic with the construction of dynamic systems that are much more elaborate than equation (3) and that are claimed to correctly represent the macro-

economic evolution. These systems, which involve many variables and equations, have been fitted to data and extensively studied.

Three remarks should be made here as to the state of the resulting unemployment macroeconomic theory:

(a) For a given economy there are usually not one but several, more or less competing, models that are supposed to represent this economy; they lead in some cases to qualitatively different predictions, even though a layman would find them rather similar to one another.

(b) The dynamic properties of these systems are complex and not even well explored; whereas stability seems to hold in most cases, diverging evolutions are also possible, which may be due to some inadequacies of the models.

(c) The models always contain equilibrium equations for phenomena in which adaptations are supposed to be fast with respect to the time unit chosen (the quarter or the year); questions must be raised as to the processes that implicitly lie behind these equations.

This theory provides, however, a favorable case in which the elaboration of the equilibrium concept was not kept separate from the study of the dynamics applying to it. The same cannot be said of other branches of economics.

(ii) A dynamic process was certainly specified for the realization of the competitive general equilibrium. It is a precise formalization of the verbal description by which L. Walras suggested how "the problem of exchange... is empirically solved". It is usually written as:

$$(4) \quad \frac{dp_h(t)}{dt} = a_h D_h[p(t)] \quad \text{for } h = 1, 2, \dots, l,$$

a_h being a positive number and $D_h(p)$ the excess-demand, which has to be zero at equilibrium according to (2):

$$(5) \quad D_h(p) = \sum_i x_{ih}(p) - \sum_j y_{jh}(p) - \omega_h.$$

But on the one hand, this process cannot claim to be an accurate representation of the law governing the actual changes of prices in industrial economies; it is still less a first approximation than the competitive equilibrium may be one. In order to describe it verbally, one has to introduce "auctioneers" who are supposed to propose prices for the various commodities and to look for equilibrium by "tâtonnement"; in actual

facts most prices are decided by the producers and wages by the employers or by collective agreements.

On the other hand, conditions for its convergence are rather limitative. They have been imposed on the matrix of derivatives [$\partial D_h(p)/\partial p_k$] and not on the more fundamental constraints and behavior that explain the formation of individual demands or supplies, $x_{ih}(p)$ and $y_{jh}(p)$. They moreover receive little support from econometric estimation.

It is fair to say that economists do not consider the process defined by (4) as anything else than an illustration of a much more elaborate mechanism in which those who fix prices use many information on their costs as well as on the market trends.²

The ideas are still less precise about the processes that may sustain equilibria in less classical theories. For instance, students of the "rational expectation equilibrium" often make reference to the learning process that leads agents to form their expectations; but this process is hardly ever spelled out; it is sometimes presented as so complex as being beyond theoretical reach. As LUCAS, Jr. (1978) wrote, "One would feel more comfortable with rational expectations equilibria if these equilibria were accompanied by some form of 'stability theory' which illuminated the forces which move an economy toward equilibrium" (p. 1429).

References

- ARROW, K., and F. HAHN, 1971, *General competitive analysis* (Holden-Day, San Francisco)
- DEBREU, G., 1979, *Existence of competitive equilibrium*, in: *Handbook of Mathematical Economics*, eds. K. Arrow and M. D. Intriligator (in print)
- KEYNES, J. M., 1936, *The general theory of employment, interest and money* (Macmillan and Co., London)
- LUCAS, Jr., R. E., 1978, *Asset prices in an exchange economy*, *Econometrica*
- MACHLUK, F., 1958, *Equilibrium and disequilibrium. Misplaced correctness and disguised politics*, *The Economic Journal*, March; reprinted in: *Essays on Economic Semantics* (Prentice-Hall Inc., 1963)
- MARSHALL, A., 1890, *Principles of economics* (Eight ed., reprinted by the Macmillan Company, New York, 1948)
- VON NEUMANN, J., and O. MORGENTERN, 1944, *Theory of games and economic behavior* (Princeton University Press)
- WALRAS, L., 1874, *Elements of pure economics*, transl. by W. Jaffé (George Allen and Urnwin, London, 1954)

² The difficulties faced by the stability analysis of the general competitive equilibrium is well recognized by ARROW and HAHN (1971) who devote three chapters to its discussion and consider other processes than the one defined by (4).

RIGHTS AND THE THEORY OF SOCIAL CHOICE

ALLAN GIBBARD

University of Michigan, Michigan, U.S.A.

I

At the core of liberalism, both in current and earlier forms, is the tenet that each person should have a protected sphere of action: that certain decisions ought, as of right, to be his alone to make, regardless of the preferences of anyone else. Among these are decisions about what to say and write and publish, what to listen to and what to read, and whether and how to worship. We know, of course, that the content of these individual liberties is problematical in various ways. Their effective exercise may require time, money, and a degree of economic independence that not all people have. One person's right to speak may conflict with another's right not to listen. A thoughtful liberal knows that any encapsulation of his position will need elucidation and refinement.

To the real problems of liberalism, social choice theory adds some bogus ones—or so it appears. Let me begin by reviewing two of these “paradoxes of liberalism”—two that together seem to me to bring out all the problems that any of them do.¹ For each person, a liberal holds, there are certain features of the world that are his business alone, and as concerns them his preferences should rule, regardless of the preferences of anyone else. That means at the very least that if two possible histories differ only in such a feature, the one he prefers is socially better. For if the history he prefers is feasible and the other obtains, then his preferences have not been the controlling ones, as they of right should have been. Call this, then, the “First Libertarian Claim”; it might be put as follows:

¹ SEN (1970) first discovered a social choice theoretic paradox of liberalism. The two given here are from GIBBARD (1974). For a bibliography and an excellent treatment of the discussion his article provoked through 1975, see SEN (1976).

Say that a person is *morally decisive* between two possible histories iff no matter what peoples' preferences were, the one he preferred would on that account be socially better. The *First Libertarian Claim* is that for each person, there are features of the world that are his business alone, and if two histories differ only in one of those features, that person is morally decisive between them. This First Libertarian Claim seems to express a part of what all liberals believe.

The claim as I have put it, by the way, is logically stronger than the claim for which SEN (1970) originally discovered a paradox. As far as I can see, though, any rationale that will support Sen's version will support mine, so that any problems for my version are problems for the rationale of Sen's.

The paradox that comes from the First Libertarian Claim is this. Suppose, to take a trivial example, the color of your bedroom walls is your business alone and the color of mine is my business alone. Suppose that I want to match your color and you want to avoid matching mine. Take a history in which both our walls are purple and one that differs from it only in that your walls are white. Since your wall color is your business alone, the First Libertarian Claim says that the one of these histories you prefer is socially better, namely your walls white and mine purple. Similar argument shows that a history that differs from this one only in that my walls too are white is better still, because I prefer it, that one that differs from that one only in that your walls are purple is even better because you prefer it, and the one with both our walls purple is better than this last one because you prefer it. The relation 'socially better than' has now cycled: Both our walls purple is better than yours purple and mine white, which is better than both white, which is better than yours white and mine purple, which is better than both purple. That is the *First Paradox of Liberalism*.

Perhaps what the paradox shows is that a preference for matching my bedroom wall color to someone else's is not the kind of preference to which my right to decide the color of my walls gives way—and likewise for your preference for distinguishing your wall color from mine. If I have the right to decide the color of my bedroom walls, the preferences which must on that account prevail are preferences genuinely for a specific color. Call such preferences "unconditional": my preference for purple walls over white is *unconditional* iff for every pair of histories that differ only in that in one my walls are purple and in the other they are white, I prefer the one in which my walls are purple. We can now formulate

a weaker *Second Libertarian Claim*, which goes like this. For each person, there are features of the world that are his business alone. If a feature is a given person's business alone and he prefers one determination of it to another unconditionally, then of any pair of histories that differ only in that the first has that feature determined in the first way and the second in the second, the first of those histories is the socially better².

The Second Libertarian Claim is inconsistent with the *Weak Pareto Principle*: that if everyone prefers one history to another, then that history is socially better. The cases that show this are much like the ones SEN discusses (1970). Assume that each person's wall color is his business alone in the sense of the Second Libertarian Claim. Suppose I want my walls to be purple, but care even more strongly that yours be purple, whereas you want your walls to be white but care even more strongly that mine be. Our preferences, then, are unconditional in the sense I have defined. Hence because of your unconditional preference for white walls, your walls white and mine purple is better than both our walls purple, and because of my unconditional preference for purple walls, both our walls purple is better than mine white and yours purple. We both, though, prefer your walls purple and mine white to your walls white and mine purple, and so by the Weak Pareto Principle, if everyone else shares that preference, the former is better. Again we have a cycle.

Now at this point we have grounds for suspicion. Whatever serious, genuine problems there may be in formulating a plausible liberalism, these do not seem to be among them. It is hard to say precisely what a person's protected sphere of action should be, but at least there are clear cases of decisions that fall within the protected sphere on any liberal view. The social choice theoretic paradoxes apply to the clear cases as fully as they do to the problematic cases. These liberal paradoxes carry with them, then, an air of sophistry: they must in some way be creating problems that do not really exist.

Whether or not studying these liberal paradoxes tells us anything about liberty, it assuredly will tell us something about social choice theory. Either the liberal paradoxes illustrate the successful application of social choice theory to problems other than voting, or there are pitfalls along the path to a widespread application of social choice theory—pitfalls that need to be understood. To talk about the paradoxes, then, is to

² This Second Libertarian Claim forms part of Hammond's proposal in this symposium, as I understand it so far.

explore the role of one kind of mathematics in thought about social norms and organization. What is it about the mathematical apparatus of social choice theory that apparently so misapplies to questions of liberty?

II

According to Robert Nozick, "The trouble stems from treating an individual's right to choose among the alternatives as the right to determine the relative ordering of these within a social ordering" (NOZICK, 1974, p. 165). Now on one interpretation of the term "social ordering" this must surely be right. When social choice theory is applied to such things as voting, giving the "social ordering" is a shorthand way of giving a rule by which canvassers are to pick an outcome: the rule the ordering stands for says to pick an alternative that, among those feasible, stands highest in that ordering. Surely if we say that I have a right to sleep on my belly or my back as I please, we do not mean that some public canvasser is to take my preference as binding when he decides in what position I shall sleep. We mean that I am to be left alone to decide for myself.

From ARROW's (1951) original work to the present, though, the mathematical apparatus of social choice theory has been given more than one interpretation. Often the "social ordering" has been interpreted not as giving a rule to be followed by public officials in picking an outcome, but as expressing an ethical judgment: as ordering the alternatives from best to worst from a moral point of view. On this construal, an Arrow social welfare function, which assigns a social ordering to every combination of individual orderings, expresses a normative theory: a theory of how the intrinsic value of a possible course of events depends on the preferences of the persons involved.³

In restating two "paradoxes of liberalism" in this paper, I have adopted this ethical interpretation of the abstract apparatus of social choice theory. What can we learn from the paradoxes on this construal? We might think, with G. E. Moore, that the concept of intrinsic value is the fundamental one in ethics, and that the content of an assertion that someone has a certain right could be explicated in terms of intrinsic value. Short of that, we might at least hold that there is an important relationship between the rights people have and the intrinsic values of possible histories.

³ SEN (1976) makes this point.

The liberal paradoxes show ways in which neither of these things can be true.

In the first place, having a private sphere of action does not entail being morally decisive over pairs of possible histories, in the technical sense of 'morally decisive' I defined: that whichever of the pair one prefers is on that account intrinsically better. For the rationale for supposing that such an entailment holds led to the first liberal paradox. If there are private spheres of action, they give rise to rights without moral decisiveness.

Perhaps the notion of having a private sphere of action is adequately captured by the Second Libertarian Claim. If it is, that tells us at least two important things. First, it tells us that the having of rights to a private sphere of action can indeed be explained in terms of intrinsic good. Second, it tells us that if more than one person has rights to a private sphere of action, then the Pareto principle is false. For on the explication provided by the Second Libertarian claim, more than one person's having rights to a private sphere of action is logically incompatible with the Pareto principle. Liberalism on this interpretation, then, commits us to the view that even if one history is preferred by everyone to another, it may not be intrinsically better.

What of a Paretian liberal? His position does not seem absurd on its face. It seems coherent to hold both that each person has the right to a private sphere of action, and that if the exercise of those rights would lead to a history which is not Pareto optimal, things are improved if people can agree on another arrangement whose outcome everyone prefers. Can this position be formulated as a condition on a social welfare function construed in the ethical way we have been discussing, as saying how the intrinsic values of possible histories depend on individual preferences?

In my article of 1974 I proposed a Pareto-consistent libertarian claim; let me now try to say what its rationale might be on the present construal of a social welfare function. For each person, the Paretian liberal thinks, there are some features of the world that are his business alone, and if two possible histories differ only in their determination of such a feature, the one that person prefers is presumptively better. The presumption settles the matter, though, only if, in a sense to be explained, it accords with a person's preferences. Say that a history z amends history w for a person i iff either the two histories differ only in a feature that is i 's business alone and i prefers z to w , or z is a Pareto improvement on w . Say that z is accessible from x without person k iff there is a sequence of

alternatives beginning with x and ending with z such that each entry in the sequence after the first amends its predecessor for someone other than k . Given an alternative x and person k , then, the alternatives accessible from x without k are the alternatives that are at least presumptively better than x by one or more applications of the Pareto principle or a right to private control on the part of someone other than k . That x be better than y *definitely accords* with k 's preferences, we might say, iff he not only prefers x to y , but also prefers to y everything else accessible from x . My *Third Libertarian Claim* says that if x and y differ only in a feature that is k 's exclusive business, then the one he prefers is better if, in this sense, its being so definitely accords with k 's preferences.⁴ This is a procedure of "look before you leap": before supposing something to be better than y just because k prefers it to y , examine all the presumptive consequences of so doing and see if they too fit k 's preferences. To suppose that y is worse than x is to suppose, at least presumptively, that y is worse than anything accessible from x ; and to do that on the basis of k 's preferences alone makes sense only if k does prefer all those alternatives to y .

It is tempting to make the requirements of liberty stronger. We might require that where x and y differ only in a feature that is k 's exclusive business and k prefers x to y , then x is better than y unless k strictly prefers y to something accessible from x . For unless k strictly prefers y to something accessible from x , the consequences of x 's being better than y accord with his preferences at least in a weak sense: that is, some agree with his strict preferences and none reverse his strict preferences. This strengthening has to be rejected not because it is implausible on its face, but because it turns out to be inconsistent. We can, however, consistently say that the mere fact that y itself is accessible from x without k will not keep k 's preference for x over y from prevailing.

The proposal as I have stated it here is slightly different from my proposal of 1974. Then I put the proposal in terms not of a social ordering but of social optimality from among feasible alternatives, and I required that accessibility be through a chain of alternatives all of which are feasible. That made presumptions arising from rights more difficult to defeat than does the version I give here. Since, though, I am now considering liberalism as a theory of intrinsic value, and since the intrinsic value of a possible history does not seem to depend on which histories are feasible, I need here to treat feasible and infeasible alternatives alike.

* For an alternative Pareto-consistent libertarian claim, see GAERTNER and KRÜGER (1980).

As I showed in effect in my article of 1974, the relations of strict social preference forced by this Third Libertarian Claim and the Weak Pareto Principle will not cycle so long as no feature is assigned to more than one person. Thus if the proposal fails as a partial theory of intrinsic good, it is not for formal reasons.

III

Most of my discussion has been of liberalism as a theory of the intrinsic value of possible histories. Liberalism seems more directly, though, to be a theory of the legitimacy of social and governmental institutions. Now the apparatus of social choice theory can be construed not only in the ways I have been discussing, but as giving a means for describing the workings of institutions. Let a *social choice function* c , for a non-empty set X of alternatives and number n of individuals, assign to each n -tuple $\langle P_1, \dots, P_n \rangle$ of preferences over X a non-empty subset of X . Let us think now of an institution as a set of rules of action; we can then use a social choice function to characterize the workings of an institution in a particular situation. Number the people who comprise the institution 1 to n , and let X be the set of histories that might result from any combination of strategies open to those people. Then we can construe $x \in c(P_1, \dots, P_n)$ as saying that x might be the outcome if each individual i had preferences P_i and each, within the constraints imposed by the institutional rules in question, rationally pursued his preferences. What the social choice function gives, then, is a dependence of outcome on individual preferences; it is the dependence that would obtain if everyone rationally pursued his preferences within the constraints of the institution.

A social choice function construed in this way can be useful both descriptively and normatively. The information it itself gives is descriptive: it tells how an institution makes outcome depend on preferences. It can, though, be put to normative use as well. For one kind of norm for an institution is a requirement on the way it makes outcome depend on preferences. Such a norm can be put as a condition on a social choice function, with the social choice function construed as saying how the institution makes outcome depend on preferences. Suppose, then, we try in this way to state the demands of liberty as a condition on a social choice function.

So understood, the First Libertarian Claim comes to this. For each person, there are features of the world that are his business alone, and

the institutions of a society are just only if they meet the following requirement. Suppose two feasible histories differ only in a feature that is the business of one person alone. Then if everyone rationally pursues his preferences subject to the constraints those institutions impose, the less preferred of those two histories will not be the outcome.

So construed, we know, the claim must be unreasonable, for for some configurations of individual preferences, it rules out every alternative. It is also clear enough why, on this construal, the First Libertarian Claim makes impossible demands on a system: No system can self-consistently decree both that my wall colors shall be the same as yours, as I wish, and that yours shall be different from mine, as you wish. Nor can any system bring about that logically impossible state of affairs indirectly, say by vesting control of each person's wall color in that person. I can only match my wall color to yours if I not only have control of my color, but know what yours will be independently of what I decide to do. But if you have the power to differentiate your color from mine whatever I decide to do, then what color your walls will be is not independent of what I decide to do, and so I cannot match you.

Can any requirement on a preferences-to-outcome function, then, succeed in expressing a consequence of liberty? It might seem that the Second Libertarian Claim does so. For suppose that, as liberty demands, the rules of my society's institutions leave me in control of a feature of the world that is my business alone. Then, it would seem, if I have an unconditional preference for one determination of that feature over another, I will certainly not determine that feature in my unconditionally less preferred way if I can determine it in my preferred way. If I prefer white walls to purple unconditionally and can paint them either way, I will not paint them purple. The outcome of a system of liberty, then, will never yield an outcome that violates the Second Libertarian Claim.

So to argue, though, presupposes that I believe that no one else will base what he does on what I do. If I do believe that, then I can treat what others do as part of an unknown state of the world which I know to be causally independent of what I am about to do. I can then apply an argument from dominance: to say that I prefer white walls to purple unconditionally is to say that I prefer white to purple whatever the state of the world may be, and in that case making my walls purple is a dominant strategy. Suppose, though, that only white and purple paint are available, and that I do indeed prefer white walls to purple for myself unconditionally. Suppose also, though, I know that if I paint mine white

you will choose to paint yours purple, whereas if I paint mine purple you will choose to paint yours white. Suppose the last thing in the world I want is for your walls to be purple. Then I will not paint my walls white, even though I unconditionally prefer them white and can paint them either color. The history that eventuates from each of us having control of his own wall color with my move first has my walls purple and yours white, even though a history which differs from it only in that both our walls are white is feasible in the relevant sense—that there are things you and I could have done that would have brought about that history. A system that gives each of us control of his own wall color violates the Second Libertarian Claim, construed as a condition on the dependence of outcome on preferences.

For an institution to satisfy the Second Libertarian Claim in this way, it must not only give each person control of the features of the world that are his business alone, but forbid anyone form basing his decisions on the ways anyone else conducts his private business. Now perhaps we do want to support such a prohibition, but it is not one that liberty is normally thought to demand, and so the kind of liberalism captured by the Second Libertarian Claim, construed as a condition on the dependence of outcome on preferences, is going to be a peculiar one at best.

IV

I think that what we learn about most from all this is some of the limitations of social choice theory as a tool for social analysis. The characteristic formal devices of social choice theory are preference aggregation functions: functions whose arguments are n -tuples of preferences over a set of alternatives, and whose values are orderings of the alternatives, sets of alternatives, choice functions on the subsets of the set of alternatives, or something of the same general kind. This class of functions is versatile in its applications: here I have discussed the use of such functions (i) to give a rule for public officials to follow in basing a public choice of some kind on the preferences of many people, (ii) to express a theory of the good which says how the intrinsic value of a possible history depends on the preferences of the people involved, and (iii) to describe a dependence of outcome on individual preferences—the dependence on institution engenders if everyone rationally pursues his preferences within the constraints of its rules. Only on the second, directly

ethical interpretation has it been possible to formulate libertarian claims that are both strong and at all plausible.

On the account of liberalism these claims provide, a liberal wants each person to control his private sphere of action because he thinks that in each case, a difference that consists of each person's getting what he wants in his private sphere of action constitutes an improvement of the world. Both the Second and the Third Libertarian Claims restrict the application of this principle in some ways, but this basic rationale remains. Now it seems to me that most liberals do not base their liberalism on any such rationale: they do not believe that each qualifying case of someone's getting what he wants is intrinsically a good thing. For a deontologist who is a liberal, this hardly needs to be said: deontologists are by definition people who think that moral rights and duties are not directly grounded in considerations of intrinsic good. What I am saying, though, applies as well to most teleologists who are liberals. John Stuart Mill's argument for liberty is that an entire system of liberty fosters the general happiness, by allowing us to learn and apply the results of the many experiments people conduct with their lives, and by allowing ideas to be tested by the clash of advocacy of the people who believe in them. None of this supposes that each time a person does what he wants in his private sphere, the mere fact that he then has what he wants in that sphere is itself an intrinsically good thing.⁵

The norms of liberalism are norms of who should be free to control what, and the tools of social choice theory—functions of arrays of preferences—have proved to be poor at representing control. It seems, then, that if we want to investigate liberty formally, we should follow the proposals of Peter GÄRDENFORS (1978) and others that we use the apparatus of game theory, namely *game forms* or “outcome functions”: functions whose arguments are arrays of individual strategies for action.⁶ I end with two diverse comments on this approach.

First, liberty is a matter of norms: to what rights people are entitled is directly a normative issue, and what rights social institutions actually accord is a matter of what norms of restraint are effectively recognized in the society. A game form itself will not represent this, as I shall explain.

⁵ Mill himself may have supposed this, for all I am claiming. The point is that Mill's utilitarian defense of liberty requires no such supposition. See *On Liberty* for the arguments of Mill to which I refer.

⁶ For an example of a problem to which social welfare functions and game forms can be jointly applied (with sad results), see GIBBARD (1978).

A game form that adequately represents a situation assigns to each person a *strategy set* consisting of all the strategies for action he can adopt, and then gives an outcome or lottery over outcomes for each *strategy profile*, or assignment to each person of a strategy in his strategy set. Now whether liberty prevails in a society is a matter of what strategies people can adopt but will not, because they regard them as illegitimate. To represent that, we need not only a game form, but a classification of strategies into those ruled out as illegitimate and those not.

Second, there will be no clearly correct way of employing this apparatus. Trying to represent rights with such an apparatus will raise all the problems that fuel controversies in less formal treatments of rights. When we say that someone has a right to control a certain feature, do we mean that it is legitimate for him to try to, that it is illegitimate to coerce him not to try to, or what? What constitutes coercion rather than mere inducement? Whether conditions placed on game forms can do much to clarify our notions of liberty is an open question, but if they can, they will do so only as a result of someone's confronting a host of difficult non-technical questions. So far, formal treatments of liberty tell us more about limitations in the applicability of certain kinds of formal apparatus than they tell us about liberty.

References

- ARROW, K. J., 1951, *Social choice and individual values*, Second edition 1963 (Wiley, New York)
- GÄRDENFORS, P., 1978, *Rights, games, and social choice*, Duplicated typescript, Department of Philosophy, University of Lund, Lund, Sweden
- GAERTNER, W., and L. KRÜGER, 1980, *Self-supporting preferences and individual rights: The possibility of the Paretian libertarianism*, *Economica*, forthcoming
- GIBBARD, A., 1974, *A Pareto-consistent libertarian claim*, *Journal of Economic Theory*, vol. 7, pp. 388-410
- GIBBARD, A., 1978, *Social choice, strategic behavior, and best outcomes*, in: *Decision Theory and Social Ethics*, eds. Gottinger and Leinfellner (Reidel, Dordrecht, Holland)
- MOORE, G. E., 1901, *Principia ethica*
- NOZICK, R., 1974, *Anarchy, state and utopia* (Basic Books, New York)
- SEN, A. K., 1970, *The impossibility of a Paretian liberal*, *Journal of Political Economy*, vol. 78, pp. 152-157
- SEN, A. K., 1970, *Liberty, unanimity, and rights*, *Economica*, vol. 43, pp. 217-245

LIBERALISM, INDEPENDENT RIGHTS AND THE PARETO PRINCIPLE*

PETER J. HAMMOND

Economics Department, Stanford University, Stanford, U.S.A.

1. Introduction

Since SEN's (1970a) paper on liberalism, economists and others have become interested in the problem of respecting individuals' rights in making social choices. Sen pointed out how easy it was for rights to come into conflict with the Pareto criterion. GIBBARD (1974) showed how one individual's rights could come into direct conflict with another's.

In fact, two rather disparate views of how individuals should be accorded rights have arisen. One view is that originally represented in Sen's work. According to this, the objectives of social choice should pay special attention to individual rights. More precisely, where an individual has a right, his preference should be decisive in determining the social welfare ordering. In other words, where choices involve matters which are the individual's own right, society accepts as legitimate the choices the individual would make according to his own preferences.

An alternative view of how individuals' rights should be respected is due especially to Buchanan. Loosely speaking, this is the doctrine of

* I am grateful to Jerzy Łoś for his careful and detailed comments, to Amartya Sen for his encouragement and patience and to him, Kotaro Suzumura, and the members of seminars at the Universities of Warwick, Stanford, Virginia and at the Virginia Polytechnic Institute—especially Kenneth Arrow, Robert Aumann, Geoffrey Brennan, Frank Hahn, Eric Maskin and John Pettengill—for their helpful comments on preliminary versions. Research support from the Social Science Research Council (U.K.) and the Institute for Mathematical Studies in the Social Sciences at Stanford University is gratefully acknowledged. A fuller version of the paper is still in preparation and will be available from the author on request.

"property rights". According to this alternative view, an individual is free to make choices of his own where he has a right to do so. Such choices may well conflict with a social welfare ordering, were one to be constructed, but that is irrelevant. Rights are presumed to be more important.

Thus, the essential difference between the two views of rights is that, according to the Sen view, there will be no conflict between the social ordering and individuals' rights, once the social ordering has been properly constructed. Whereas, with the property rights view, there may well be such a conflict, but this conflict is always resolved by giving individual rights precedence in the social choice procedure.

A second difference between these two views is that, in the Sen view, the final social choice will always maximize a social welfare ordering subject only to feasibility constraints. With property rights, however, individuals' private choices impose further restrictions on the social choice, and so, in general, the final social choice will not maximize an ordering in the usual way.

In fact, the social choice problem with property rights is very reminiscent of the optimal taxation problem in public finance theory. With optimal taxes, each individual is free to maximize his own preference ordering over a budget constraint which is determined by the taxation authorities. The taxation authorities then determine the budget constraints so that the resulting economic allocation maximizes a social welfare ordering, subject to the constraint that individuals choose what they want within their budget constraints. With property rights, the social welfare ordering determines the social state subject to the constraints imposed by the way individuals choose to exercise their property rights.

Indeed, it turns out for our purposes that the main difference between the Sen view of individual rights on the one hand, and property rights on the other, is whether one forces social choices to be ordinal or not. Sen does force them to be ordinal, by constructing a social ordering which respects rights. With property rights, certain options are vetoed by individuals and deleted from the social choice set, but there is no attempt to force social choices to conform to an ordering.

Both kinds of rights lead rather easily to logical difficulties, and it is notable that the two rather different kinds of rights experience very similar logical difficulties, whose resolution is also very similar. Thus, we have the "Gibbard paradox" (GIBBARD, 1974) which arises when, in attempting to give individuals their rights in the sense of Sen, cycles are inevitably introduced into the social welfare strict preference relation. In this case

too, if one tries to give individuals "property rights", there is no determinate Nash equilibrium in pure strategies.

A second difficulty with both ways of respecting rights is that the outcome of the social choice procedure can easily violate the Pareto criterion, in that it can select Pareto inefficient outcomes. This is the original "Sen paradox" (SEN, 1970a).

The present paper, which follows many others on the subject, is an attempt to establish fairly generally exactly what sort of rights can be respected by the social choice rule, and how far the Pareto criterion can be followed without violating individuals' rights. To overcome the "Gibbard paradox", I show how to construct "privately unconditional" restricted preferences according to a certain dominance relation and how it is always possible to respect rights in the limited sense that, where an individual has a right to decide an issue, then his privately unconditional restricted preferences will decide it. I also show how to construct "privately oriented" preferences according to another, rather stronger, dominance relation and show that it is always possible to respect rights as regards privately unconditional preferences and simultaneously to satisfy the Pareto criterion as regards privately oriented preferences. This circumvents the "Sen paradox".

In defining privately unconditional and privately oriented restricted preferences, I shall assume throughout the present abbreviated version of the paper that rights are assigned independently, in the sense that there is a product space with a public component and, for each individual, a private component. Then the definition of the privately oriented preference dominance relation accords quite closely with GAERTNER and KRÜGER (1980) concept of "self-supporting" preferences. Independence of rights simply means that each individual is decisive over his own private component. Conditions to ensure that rights are independent will be presented in the full version of the paper.

This abbreviated version considers just individual rights, as does nearly all the existing literature—BATRA and PATTANAIK (1972) is an early exception, with its concept of "minimal federalism". The later full version will extend to group rights most of the results which apply to individual rights.

Section 2 of the present paper defines social choice rules, discusses the limitations imposed by neutrality and explains what is meant by a "rights profile." Section 3 introduces independent rights and shows how respecting rights as regards privately unconditional preferences is always possible.

Section 4 discusses privately oriented preferences and shows how a "private" Pareto criterion never conflicts with individual rights as regards their privately unconditional preferences. Section 5 contains some discussion of the results and conclusions.

2. Social choice rules and individual rights

2.1. *Social choice rules.* Let X denote the fixed underlying set of social options. Let N denote the finite set of individual members of the society. Assume that each individual $i \in N$ has a preference ordering R_i on X (a preference ordering is a reflexive, complete and transitive binary relation). " $aR_i b$ " is to be interpreted as " a is no worse than b ". Let $\underline{R} = (R_i)_{i \in N}$ denote a *profile* of preference orderings, one ordering for each individual member of the society.

A *strict preference relation* on X is an irreflexive but not necessarily transitive binary relation which will usually be denoted by P .

Given the profile \underline{R} , each individual $i \in N$ has a transitive strict preference relation P_i given by:

$$aP_i b \Leftrightarrow \text{not } bR_i a.$$

Here, " $aP_i b$ " can be interpreted as " a is preferred to b by i " or as " a is better than b for i ".

Let $\mathcal{R}(X)$ denote the set of all logically possible preference orderings on the set X . Let $\mathcal{R}^N(X)$ denote the set of all logically possible profiles of preference orderings \underline{R} . Thus, by definition:

$$\underline{R} \in \mathcal{R}^N(X) \Leftrightarrow \text{for all } i \in N: R_i \in \mathcal{R}(X).$$

Let $\mathcal{P}(X)$ denote the power set consisting of all subsets of the underlying set X . Then a social choice rule is a mapping $C: \mathcal{P}(X) \times \mathcal{R}^N(X) \rightarrow \mathcal{P}(X)$ with the following properties:

- (1) For every $A \subseteq X$ and every $\underline{R} \in \mathcal{R}^N(X)$, $C(A, \underline{R}) \subseteq A$.
- (2) For every finite $A \subseteq X$ and every $\underline{R} \in \mathcal{R}^N(X)$, $C(A, \underline{R})$ is non-empty.

Thus, for any fixed profile \underline{R} , $C(\cdot, \underline{R})$ is a (social) choice function on the underlying set X (cf. HERZBERGER, 1973). An important special case which underlies the Sen approach to liberalism occurs when the choice function $C(\cdot, \underline{R})$ is ordinal, and corresponds to a social welfare ordering $R = f(\underline{R})$ which is function of the profile \underline{R} . In other words:

$$C(A, \underline{R}) = \{a \in A \mid af(\underline{R})b \text{ for all } b \in A\}.$$

Then the mapping f is an Arrow social welfare function, now to be defined formally.

An *Arrow social welfare function* f (or “constitution”) is a mapping, defined on a domain $\mathcal{R} \subseteq \mathcal{R}^N(X)$, such that $f(\tilde{R}) \in \mathcal{R}(X)$ for all $\tilde{R} \in \mathcal{R}^N(x)$. Thus, f determines the social welfare ordering $R = f(\tilde{R})$ as a function of the preference profile \tilde{R} . It is possible to allow R to depend on things other than just the profile \tilde{R} —for instance, interpersonal comparisons of utility, the status quo, etc.—but this would merely add inessential complications to the notation.

In the rest of this paper, I shall use “ASWF” to refer to an Arrow social welfare function.

2.2. Individual rights relations. As discussed in the introduction, I am going to follow SEN (1970a, 1970b) and others in interpreting a “right” as meaning that an individual’s own preferences over certain “personal” or “private” matters are to be reflected in the social choice rule. Formally, each individual $i \in N$ will have his rights described by a binary decisiveness relation D_i on the set X . If individual i ’s rights are respected in full in an ASWF, this will mean that $aD_i b$ and $aP_i b$ together imply aPb (where P denotes the social strict preference relation corresponding to the social ordering $R = f(\tilde{R})$). On the other hand, if each individual i ’s property rights are fully respected in the social choice rule, this will mean that if $b \in A$ and there exist $i \in N$ and $a \in A$ such that $aD_i b$ and $aP_i b$, then $b \notin C(A, R)$. In other words, with property rights, each individual can *veto* an outcome b if he prefers an alternative outcome a and he has a right to choose between a and b .

In general, as Sen and GIBBARD (1974) have already pointed out, it is not always possible to respect each individual’s rights fully: nonetheless, for the moment I shall allow D_i to be a fairly general relation and then later explore the extent to which rights described by the profile of relations $(D_i)_{i \in N}$ can be respected.

In fact, I shall make the following assumption:

A.1: For each $i \in N$, the (binary) *rights relation* D_i is reflexive, symmetric and transitive on X : D_i may, however, just be the diagonal relation Δ given by:

$$a\Delta b \Leftrightarrow a = b \text{ (all } a, b \in X\text{)}.$$

Of course, where $D_i = \Delta$, individual i has no rights. In fact, reflexivity of D_i has no significance beyond that of making D_i an equivalence relation.

The conditions of symmetry and transitivity are plausible in many cases but they can be excessively restrictive. In particular, symmetry may be unappealing where matters of life and death are involved.

Together with transitivity, symmetry implies that, for each i , we can partition the space X into D_i -equivalence classes $[x]_{D_i}$ such that

$$a \in [x]_{D_i} \Leftrightarrow a D_i x.$$

Then i 's own preference relation, where it is strict, will be decisive in selecting a social option from the equivalence class $[x]_{D_i}$; however, choices between elements of different equivalence classes $[x]_{D_i}, [x']_{D_i}$ will depend on other individuals' preferences in general.

A *rights profile* \underline{D} is a list of rights relations $(D_i)_{i \in N}$, one for each individual.

Given the social rule $C: \mathcal{P}(X) \times \mathcal{R}^N(X) \rightarrow \mathcal{P}(X)$ and the rights profile \underline{D} , say that C respects \underline{D} fully if, wherever $b \in A$ and there exist $a \in A$ and $i \in N$ such that $a P_i b$ and $a D_i b$, then $b \notin C(A, R)$. If there is an Arrow social welfare function, it will also be said to respect rights fully if the associated social choice rule does.

3. Independent rights and privately unconditional preferences

3.1. *Independent rights.* I shall describe two different versions of "independent rights". Each will involve the underlying set X being a subset of a product space:

$$X \subseteq Y := Y_N \times \prod_{i \in N} Y_i$$

where each $i \in N$ has a component space Y_i and the group also has a component space Y_N . Economists may find it helpful to regard $y_i \in Y_i$ as i 's personal consumption vector, and $y_N \in Y_N$ as a vector of public goods.

Given the product space $Y = Y_N \times \prod_{i \in N} Y_i$, say that $(y = y_N, (y_i)_{i \in N})$ and $y' = (y'_N, (y'_i)_{i \in N})$ are i -variants if $y_N = y'_N$ and $y_j = y'_j$ (all $j \in N - \{i\}$). Write y_{-i} for $(y_N, (y_j)_{j \in N - \{i\}})$, so that y and y' are i -variants if and only if $y_{-i} = y'_{-i}$. Then rights are said to be *weak independent* if:

- (i) $X \subseteq Y_N \times \prod_{i \in N} Y_i$.
- (ii) For all $i \in N$, $y D_i y' \Rightarrow y$ and y' are i -variants.

Rights are said to be *strong independent* if:

- (i) $X \subseteq Y_N \times \prod_{i \in N} Y_i$.
- (ii) For all $i \in N$, $y D_i y' \Leftrightarrow y$ and y' are i -variants.

Thus, with weak independent rights, i may not be decisive over certain issues affecting just his component of the product space, whereas with strong independent rights he must be.

3.2. Gibbard's example. The following example is due to GIBBARD (1974) and is beautifully illustrated in SEN (1976). Let $N = \{1, 2\}$. Take $X = Y_1 \times Y_2$ where $Y_1 = \{a_1, b_1\}$ and $Y_2 = \{a_2, b_2\}$. Suppose that rights are strong independent. Let $f: \mathcal{R} \rightarrow \mathcal{R}$ be a social welfare function with the domain \mathcal{R} unrestricted. Then the following profile R shows that f cannot respect rights fully:

$$(a_1, a_2)P_1(b_1, a_2)P_1(b_1, b_2)P_1(a_1, b_2), \\ (a_1, b_2)P_2(a_1, a_2)P_2(b_1, a_2)P_2(b_1, b_2).$$

(If a_1, a_2 are in some sense alike, and so are b_1, b_2 , then individual 1 can be regarded as a conformist and individual 2 as a nonconformist.) If rights were respected fully in an ASWF, one would have to have:

$$(a_1, a_2)P(b_1, a_2)P(b_1, b_2)P(a_1, b_2)P(a_1, a_2)$$

which is a contradiction. Similarly, if rights are respected fully by a social choice rule, then $C(X, R)$ would have to be empty—again a contradiction.

The trouble in this example arises from the attempt to respect rights fully over too large a domain \mathcal{R} of preference profiles. A resolution can be achieved by restricting the domain, or by giving limited rights. In fact, I shall prefer to limit individuals' rights, rather than to restrict the domain of profiles as, for instance, BREYER (1977) does. My approach will turn out to be more general since, in the special case where the profile lies in a domain which is sufficiently restricted to resolve this kind of paradox then the profile is such that rights can be respected fully. On the other hand, if the profile lies outside such a domain, then an approach to liberalism which insists on such domain restrictions before individual rights can be respected loses all its force. My approach still respects individual rights as far as can be done without a contradiction, and is more general than that of GAERTNER and KRÜGER (1979).

In what follows I shall show how to construct "privately unconditional" restricted preferences such that individuals' rights can be respected in the sense that their privately unconditional preferences are decisive in determining the social choice, even though their original unrestricted preferences may not be (cf. GIBBARD, 1974). These "privately unconditional" restricted preferences will be based on a dominance relation.

In Example 3.2, it will turn out that the privately unconditional restricted preferences are null for both individuals, so that *any* social welfare function respects rights in the limited sense I shall be using.

3.3. Privately unconditional preferences. Suppose that rights are strong independent and let i be any individual of the set N . Say that \bar{P}_i is a *privately unconditional* strict preference relation if, for every fixed $\bar{y}_N \in Y_N$, there exists an irreflexive and transitive strict preference relation $P_i^0(\bar{y}_N)$ on Y_i such that, for any i -variants $a, b \in X$ satisfying $a_{-i} = b_{-i}$ and $a_N = b_N = \bar{y}_N$:

$$a\bar{P}_i b \Leftrightarrow a_i P_i^0(\bar{y}_N) b_i.$$

So a privately unconditional strict preference relation \bar{P}_i applied to issues affecting i 's personal component of the product space Y_i , must be equivalent to a relation $P_i^0(\bar{y}_N)$ on Y_i .

For any preference ordering R_i , and any fixed $y_N \in Y_N$, define the irreflexive and transitive strict preference relation $P_i^0(R_i, \bar{y}_N)$ on Y_i so that $y_i P_i^0(R_i, \bar{y}_N) y'_i$ if and only if $a P_i b$ for every pair of i -variants $a, b \in X$ satisfying $a_i = y_i$, $b_i = y'_i$ and $a_N = b_N = \bar{y}_N$. Thus $P_i^0(R_i, \bar{y}_N)$ is a conditional dominance relation on the set Y_i . Then define the *privately unconditional restriction* $\bar{P}_i(R_i)$ of P_i so that $a\bar{P}_i(R_i)b$ if and only if a and b are i -variants, $a_N = b_N = \bar{y}_N$, and $a_i P_i^0(R_i, \bar{y}_N) b_i$.

Thus $\bar{P}_i(R_i)$ is the strongest possible privately unconditional strict preference relation defined on pairs of i -variants which is restriction of i 's (possibly privately conditional) strict preference relation P_i . Notice that $\bar{P}_i(R_i)$ is irreflexive and transitive. In fact, $a\bar{P}_i(R_i)b$ if and only if a and b are i -variants and a_i dominates b_i for individual i , in an obvious sense.

Because rights are strong independent notice that $a\bar{P}_i(R_i)b$ implies that aD_ib .

3.4. Respecting rights conditionally. Say that the social choice rule $C(A, R)$ respects the rights profile D conditionally provided that, whenever $b \in A$ and there exist $i \in N$ and $a \in A$ such that $a\bar{P}_i(R_i)b$ (where $\bar{P}_i(R_i)$ is the privately unconditional restriction of i 's strict preference relation), then $b \notin C(A, R)$. If there is an ASWF, it will be said to respect rights conditionally provided that the associated social choice rule does.

Following GIBBARD (1974), it can now be shown that, provided rights are strong independent, it is always possible to respect them in this limited sense. But Gibbard's Example 3.2 above shows how, even if rights are

strong independent, it is not possible to respect them fully over an unrestricted domain of preference profiles.

THEOREM 3.4 (cf. GIBBARD, 1974). *Suppose that rights are strong independent. Then there exists a social choice rule and, indeed, a social welfare function, which respects these rights conditionally.*

PROOF: This result is a special case of Theorem 4.3 below and so the proof is omitted.

4. Privately oriented preferences and the private Pareto principle

4.1. *Sen's Example.* (SEN, 1970a, 1970b.) Let $N = \{1, 2\}$. Take $X = Y = Y_1 \times Y_2$, where $Y_1 = \{a_1, b_1\}$, $Y_2 = \{a_2, b_2\}$. Suppose that rights are strong independent. Let $f: \bar{\mathcal{R}} \rightarrow \mathcal{R}$ be a social welfare function with the domain $\bar{\mathcal{R}}$ of privately unconditional preferences. Consider the following profile (of privately unconditional preferences):

$$\begin{aligned} & (b_1, b_2)P_1(a_1, b_2)P_1(b_1, a_2)P_1(a_1, a_2), \\ & (a_1, a_2)P_2(a_1, b_2)P_2(b_1, a_2)P_2(b_1, b_2). \end{aligned}$$

(Sen interprets a_i as meaning that i reads *Lady Chatterley's Lover*, and b_i as meaning that i does not read it. Person 1 is a "prude" and person 2 is "lascivious". I have introduced the extra option (a_1, a_2) to show that the product structure of the space can be complete—i.e. rights can be strong independent—without destroying the example.)

If f respects rights, conditionally or fully, then the following must be satisfied by social preferences:

$$(b_1, a_2)P(b_1, b_2)P(a_1, b_2); \quad (b_1, a_2)P(a_1, a_2)P(a_1, b_2).$$

In any case, $(b_1, a_2)P(a_1, b_2)$, although both individuals strictly prefer (a_1, b_2) to (b_1, a_2) . Thus, we see a conflict between respecting rights ("liberalism") and the Pareto principle. The trouble here is the attempt to apply the Pareto rule to non-privately oriented preferences. In fact, according to the definitions in Section 4.2 below the privately oriented preference relations \hat{P}_1 and \hat{P}_2 must satisfy:

$$\begin{aligned} & (b_1, b_2)\hat{P}_1(a_1, a_2); \quad (b_1, b_2)\hat{P}_1(a_1, b_2); \quad (b_1, a_2)\hat{P}_1(a_1, a_2); \\ & (b_1, a_2)\hat{P}_1(a_1, b_2); \quad (a_1, a_2)\hat{P}_2(b_1, b_2); \quad (a_1, a_2)\hat{P}_2(a_1, b_2); \\ & (b_1, a_2)\hat{P}_2(b_1, b_2); \quad (b_1, a_2)\hat{P}_2(a_1, b_2) \end{aligned}$$

and there is no conflict between respecting rights and the Pareto principle applied to these privately oriented preferences. Respecting rights implies that the social preference relation must satisfy:

$$\begin{aligned} (b_1, a_2)P(a_1, a_2)P(a_1, b_2), \\ (b_1, a_2)P(b_1, b_2)P(a_1, b_2) \end{aligned}$$

but is otherwise unrestricted. The chosen outcome, of course, is (b_1, a_2) .

4.2. Privately oriented preferences. Suppose that rights are strong independent. Let i be any individual of the set N .

Say that \hat{P}_i is a *privately oriented* strict preference relation for i if there exists a strict preference relation \hat{P}_i^0 on $Y_N \times Y_i$ such that, for every $a, b \in X$:

$$a\hat{P}_i b \Leftrightarrow (a_N, a_i)\hat{P}_i^0(b_N, b_i).$$

So \hat{P}_i is privately oriented if it ignores the private issues of individuals other than i , and is equivalent to a strict preference relation \hat{P}_i^0 on just the product space $Y_N \times Y_i$ of public issues and i 's private issues.

For any preference ordering R_i for individual i , define the irreflexive and transitive strict preference relation $\hat{P}_i^0(R_i)$ on $Y_N \times Y_i$ so that $(y_N, y_i)\hat{P}_i^0(R_i)(y'_N, y'_i)$ if and only if $aP_i b$ for every pair $a, b \in X$ satisfying $a_j = b_j$ (all $j \in N - \{i\}$), $(a_N, a_i) = (y_N, y_i)$ and $(b_N, b_i) = (y'_N, y'_i)$. Thus $\hat{P}_i^0(R_i)$ is a dominance relation on the set $Y_N \times Y_i$. Then define i 's *privately oriented preferences* $\hat{P}_i(R_i)$ so that $a\hat{P}_i(R_i)b$ if and only if $(a_N, a_i)\hat{P}_i^0(R_i)(b_N, b_i)$.

Thus $\hat{P}_i(R_i)$ is the privately oriented preference relation which is equivalent to the dominance relation $\hat{P}_i^0(R_i)$ on $Y_N \times Y_i$. Notice that $\hat{P}_i(R_i)$ is irreflexive and transitive. Of course, it need not be a restriction of P_i : in Sen's Example 4.1 above, $(b_1, a_2)\hat{P}_1(R_1)(a_1, b_2)$ and $(b_1, a_2)\hat{P}_2(R_2)(a_1, b_2)$ although $(a_1, b_2)P_1(b_1, a_2)$ and $(a_1, b_2)P_2(b_1, a_2)$.

Evidently, when rights are strong independent, if P_i is privately oriented it must be privately unconditional with $P_i^0(\bar{y}_N)$ defined by:

$$a_i P_i^0(\bar{y}_N) b_i \Leftrightarrow (\bar{y}_N, a_i) \hat{P}_i^0(\bar{y}_N, b_i).$$

But a privately unconditional preference relation need not be privately oriented, as is clear from Sen's Example 4.1 above. Nevertheless:

LEMMA 4.2. *Suppose that rights are independent. If $\bar{P}_i(R_i)$ is the privately unconditional restriction of P_i and if $\hat{P}_i(R_i)$ are i 's privately oriented preferences, then $a\bar{P}_i(R_i)b$ if and only if a and b are i -variants and $a\hat{P}_i(R_i)b$.*

PROOF: Define $\bar{y}_N := a_N = b_N$ when a, b are i -variants. Then:

$$\begin{aligned} a\bar{P}_i(R_i)b &\Leftrightarrow a, b \text{ are } i\text{-variants and } a_iP_i^0(\bar{y}_N, R_i)b_i \\ &\Leftrightarrow a, b \text{ are } i\text{-variants and } (\bar{y}_N, a_i, y_{-i})P_i(\bar{y}_N, b_i, y_{-i}) \\ &\quad (\text{all } y_{-i} \in Y_{-i}) \\ &\Leftrightarrow a, b \text{ are } i\text{-variants and } (\bar{y}_N, a_i)\hat{P}_i^0(R_i)(\bar{y}_N, b_i) \\ &\Leftrightarrow a, b \text{ are } i\text{-variants and } a\hat{P}_i(R_i)b. \quad \blacksquare \end{aligned}$$

4.3. The private Pareto principle. The private Pareto principle is obtained by applying the ordinary Pareto principle to the individuals' privately oriented preferences. Thus, the social choice rule $C(A, \underline{R})$ is privately Paretian if, whenever $b \in A$ and there exists $a \in A$ such that $a\hat{P}_i(R_i)b$ for all $i \in N$, then $b \notin C(A, \underline{R})$. Here, of course, $\hat{P}_i(R_i)$ is i 's privately oriented preference relation, given the preference ordering R_i and the strong independent rights profile \underline{D} .

THEOREM 4.3. *Suppose that rights are strong independent. Then there exists a social choice rule and, indeed, a social welfare function, which is privately Paretian and also respects these rights conditionally.*

PROOF: (1) Define the irreflexive social strict preference relation P as follows: aPb if

- (i) there exists $i \in N$ such that $a\bar{P}_i(R_i)b$ (where $\bar{P}_i(R_i)$ is the privately unconditional restriction of i 's preference ordering)
or
- (ii) for every $i \in N$, $a\hat{P}_i(R_i)b$.

(2) It suffices to prove that the relation P is acyclic because then:

- (i) There exists a social choice rule $C(A, \underline{R})$ (with $C(A, \underline{R})$ non-empty whenever A is finite) such that $x \in C(A, \underline{R})$ only if $x \in A$ and there is no $y \in A$ for which yPx . By construction $C(A, \underline{R})$ respects rights conditionally.
- (ii) The relation P has an irreflexive and transitive extension P^* defined by: $xP^*y \Leftrightarrow \exists a^1, a^2, \dots, a^n: xPa^1, a^1Pa^2, \dots, a^{n-1}Pa^n, a^nPy$. There exists an ASWF $f(\underline{R})$ in which for every profile \underline{R} , the ordering $R = f(\underline{R})$ is an order-extension of the relation P^* . Such an order extension exists as a consequence of Szpilrajn-Marczewski theorem (see, e.g. SEN, 1970b). Moreover, f will then respect rights conditionally.

(3) Suppose that P is not acyclic. Then there exists a finite set $A = \{a^1, a^2, \dots, a^n\}$ such that:

$$a^1Pa^2, a^2Pa^3, \dots, a^{r-1}Pa^r, \dots, a^{n-1}Pa^n \text{ and } a^nPa^1.$$

I shall show that this leads to a contradiction by constructing a real-valued function w on the domain A with the property that if xPy and $x, y \in A$ then $w(x) > w(y)$. This will then prove the theorem.

(4) To construct the function w , first notice that because A is finite, so are the projections A_i of the set A onto Y_i (all $i \in N$), and the projection A_N of A onto Y_N , defined by:

$$A_i := \{a_i \mid \exists y_N \in Y_N, \exists y_{-i} \in Y_{-i}: (y_N, a_i, y_{-i}) \in A\},$$

$$A_N := \{a_N \mid \exists y_i \in Y_i \text{ (all } i \in N\text{)}: (a_N, (y_i)_{i \in N}) \in A\}.$$

(5) Because $A_N \times A_i$ is finite, for each $i \in N$, we can construct “one-way representations” v_i of $\hat{P}_i^0(R_i)$ on $A_N \times A_i$. In other words, for each $i \in N$, we can construct a real-valued function v_i on $A_N \times A_i$ such that

$$(a_N, a_i)\hat{P}_i^0(R_i)(b_N, b_i) \text{ implies } v_i(a_N, a_i) > v_i(b_N, b_i).$$

Define the real-valued function w on A by:

$$w(y) := \sum_{i \in N} v_i(y_N, y_i).$$

Now suppose aPb where P is defined as in part (1) of this proof. Two cases are possible:

- (i) There exists a unique $i \in N$ such that $aD_i b$ and also $a\bar{P}_i(R_i)b$. But then, by Lemma 4.2, a, b are i -variants and $a\hat{P}_i(R_i)b$; in fact, $(a_N, a_i)\hat{P}_i^0(R_i)(b_N, b_i)$. Therefore $v_i(a_N, a_i) > v_i(b_N, b_i)$ by construction of v_i . Because a, b are i -variants, $(a_N, a_j) = (b_N, b_j)$ for all $j \in N - \{i\}$. Therefore $w(a) > w(b)$.
- (ii) For every $i \in N$, $a\hat{P}_i(R_i)b$. But then, for every $i \in N$, $(a_N, a_i)\hat{P}_i^0(R_i)(b_N, b_i)$ and so $v_i(a_N, a_i) > v_i(b_N, b_i)$. It follows that $w(a) > w(b)$.

So, in either case, aPb implies $w(a) > w(b)$, as required. ■

The assumption of strong independence plays a crucial role in this proof by allowing the use of Lemma 4.2, and by allowing $aD_i b$ to imply that a and b must be i -variants.

5. Discussion and conclusions

In this paper, I have explored how far rights can be respected in a social choice rule, be they property rights or the “decisive” rights which underlie Sen’s concept of liberalism. To avoid contradictions, individuals cannot generally be given more than conditional rights, so that individual preferences are only decisive in determining the social outcome where they are “privately unconditional”. And to avoid contradicting the Pareto principle, this principle has to be modified so that it is only applied where individuals’ preferences are “privately oriented”, which amounts to ruling out external diseconomies. These restrictions on rights and on the scope of the Pareto principle are far from trivial and may be regarded as too strong, particularly where Sen’s “decisive” rights are under consideration. It may be thought desirable to heed an individual’s rights even though his preferences are privately conditional, or to heed an individual’s preferences in applying the Pareto principle even though his preferences are not privately oriented. But then something else has to give; somebody’s rights may have to be violated even though their preferences are unconditional; or somebody’s preferences may have to be ignored in applying the Pareto principle, even though their preferences are privately oriented. Neither is satisfactory.

Where property rights are concerned, and these rights are strongly independent, it is as if the individuals in the society were involved in a game form, with their personal choices as strategies. Generally, to have a determinate outcome, each individual must always have a dominant strategy, which will only occur if preferences are privately unconditional. And this determinate outcome will only be Pareto efficient in general if preferences are privately oriented. Where preferences are not privately unconditional, a Nash equilibrium may not exist and so there may be no determinate outcome. Instead, individuals will choose a probability mixture of undominated strategies, in general, and the outcome will be a probability mixture of outcomes in the social choice set $C(A, R)$. Each of the outcomes in the probability mixture involves undominated strategies for each individual, which means precisely that their rights are respected conditionally. Moreover, each outcome in the probability mixture is Pareto efficient in the limited sense of the private Pareto principle.

Unsatisfactory as this is, it is simply not possible to go further in respecting individual rights, possibly in combination with the collective

Pareto principle, without examining carefully the degree of agreement between different individuals' preferences, rather than just looking separately at each individuals' preferences as this paper has done.

References

- BARRY, B. M., 1965, *Political argument* (Routledge and Kegan Paul: London)
- BATRA, R. N., and P. K. PATTANAIK, 1972, *On some suggestions for having non-binary social choice functions*, Theory and Decision, vol. 3, pp. 1-11
- BLAU, J. H., 1975, *Liberal values and independence*, Review of Economic Studies, vol. 42, pp. 395-401
- BREYER, F., 1977, *The liberal paradox, decisiveness over issues, and domain restrictions*, Zeitschrift für Nationalökonomie, vol. 37, pp. 45-60
- BUCHANAN, J. M., 1975, *The limits of liberty: between anarchy and leviathan*
- BUCHANAN, J. M., 1976, *The justice of natural liberty*, Journal of Legal Studies, vol. 5(1), pp. 1-16
- FARRELL, M. J., 1976, *Liberalism in the theory of social choice*, Review of Economic Studies, vol. 43, pp. 3-10
- GAERTNER, W., and L. KRÜGER, 1980, *Self-supporting preferences and individual rights: The possibility of Paretian libertarianism*, Economica (forthcoming)
- GIBBARD, A., 1974, *A Pareto-consistent libertarian claim*, Journal of Economic Theory, vol. 7, pp. 399-410
- HERZBERGER, H., 1973, *Ordinal preferences and rational choice*, Econometrica, vol. 41, pp. 187-218
- SEIDL, C., 1975, *On liberal values*, Zeitschrift für Nationalökonomie, vol. 35, pp. 257-292
- SEN, A. K., 1970a, *The impossibility of a Paretian liberal*, Journal of Political Economy, vol. 78, pp. 152-157
- SEN, A. K., 1970b, *Collective choice and social welfare* (Holden-Day, San Francisco)
- SEN, A. K., 1976, *Liberty, unanimity and rights*, Economica, vol. 43, pp. 217-246
- SUZUMURA, K., 1978, *On the consistency of libertarian claims*, Review of Economic Studies, vol. 45, pp. 329-342

MATHEMATICAL ANALYSIS OF LANGUAGE

ZELLIG HARRIS

It is possible to develop a theory of language in which mathematics plays a role different from its use in most of natural science. This difference is not because language is a human and purposive product rather than a part of objectively given nature. Indeed, language does not have to be studied initially as a human endeavor, with what tools are available for such studies. Instead, one can first look upon language as an aggregate of occurrences of speaking and writing. There are reasons for doing so. Since the intent of the speaker and the effect of speaking can be studied only imprecisely, the possibilities for precise analysis lie rather in the physical events of speaking and writing, as combinations of sounds and of letters, taken in the first instance as events which are characterized not by their relation to the human world but purely by their internal structure.

The problem then is to characterize the events of speaking and writing, that is, to state the parts and their combinations, or the elements and their relations, or other properties that hold for all such events and only for them. Some considerations seem clear from the start. The fact of alphabetic writing shows that discrete objects, the letters of the alphabet, suffice for a representation of language, excluding unspecified expressive inflections. And the experimental method called 'phonemic analysis' shows that the continuous flow of sound in each word can be represented as a succession of discrete phonemes, with just a small set of phonemes sufficing to distinguish the many sounds which occur in speaking a language.

The first picture that one obtains of the structure of these events is that they can each be segmented into successive sentences, with each sentence representable as a sequence of words, or of stems with affixes (all of these called 'morphemes'), and each word or morpheme representable

as a sequence of phonemes. This was the framework of structural linguistics. The next step was to state regularities in the word-successions that constituted sentences (as against those that did not), recognizing for example that *the man walked*, *the man came*, *the man left* are sentences while **the man hotel*, **the man universe* are not. Each of these regularities holds over a particular domain of words, and the domains of many regularities are coextensive or almost so. To take a simple case, the words that appear before *is here*, *is missing*, etc., to make a sentence include *the pen*, *the light*, *the fork*, *the knife* (but not *the knive*), while the words that appear before *-s are here*, *-s are missing*, etc., include *the pen*, *the light*, *the fork*, *the knive*, but not *the knife*. The two domains can be made identical, if we consider *knive* to be a variant form of *knife*, the occurrence of the variant being determined by this *-s*; then the same list occurs before *is here* and *-s are here*. In another type of case, we have *work*, *think*, *sing* and many other words all appearing after *I*, *you*, *we*, *the children* and many other words, to make a sentence, but *am* only after *I* (in *I am*) and *are* only after the other words (in *You are*, *We are*, *The children are*). Here again, the domains can be made identical if we take *am* as the post-*I* variant of *are*, so that the pair *am*, *are* form a single member of the domain which includes *work*, *think*, *sing*, and which forms a sentence after *I*, *you*, etc. Such examples are best stated if we deal, for the time being, only with very short sentences.

When such regularizations of domain are carried out to the fullest extent permitted by the actually occurring combinations of words, we reach a structural description in which a great many of the exceptions so characteristic of grammar are eliminated. More exactly, these exceptions are transferred from creating restricted domains to creating variant-pairs within regularized domains.

Given this structural picture, another method, called 'transformational analysis', provides a further simplification in the formulation of how word-successions make sentences. This analysis arises as follows: In the structural description referred to above, we can say, for example, that a sentence results if a word of one set *A* (a domain of many regularities) is followed by a word of another set *B*. Thus, a word from *I*, *children*, *motors*, etc., followed by a word from *work*, *sing*, *think*, etc., makes a sentence. However, not all combinations are equally likely to occur: *I work*, *I think*, etc., are reasonably likely to be said, as is *Motors work*, and less so *Motors sing*, but hardly *Motors think*. Grammars never tried to specify the individual word-combinations that are reasonably likely to

occur in sentences, as against those combinations that are not, because the data was far too complex and shifting, especially for long sentences. But the problem can be reduced, yielding by the way a new picture of the structure of sentences. This is done as follows:

Consider first the question of long sentences. If we look at the word-combinations in a long sentence, we often find that the sentence can be segmented into parts each of which contains the word-combinations of short sentences. For example, in short sentences, *the book* occurs before *fell*, *was lost*, *cost \$ 5*, *interested me*, (or after *I found*, *He bought*, etc.) but hardly ever before *slept*, *drank wine*, *coughed*. In longer sentences, containing *the book which*, the words immediately after *which* are from the first set and not the second: *The book which was lost cost \$ 5*, *The book which cost \$ 5 was lost*, *I found the book which was lost*, *The book which I found interested me*, but not **The book which coughed cost \$ 5*. Indeed, we find in these sentences two occurrences of the words that can occur with *book*. We then say that these sentences are each formed out of two shorter ones each containing *book*; *The book cost \$ 5*; *the book was lost* with a change of the repeated noun (*the book*) to *which*, and moving of the second sentence to after the first occurrence of the repeated noun (the antecedent of *which*). This provides a new analysis of *The book which was lost cost \$ 5*, not as a sentence containing a relative clause, but as a sequence of two short sentences of which the second underwent certain changes. The concept of 'relative clause' can thus be eliminated from grammar.

The similarities of word-combination can be found not only in different segments of a single long sentence, but also as between two structurally different short sentences. Thus if we consider verbs that appear both as transitives (with a noun object: *He reads poetry*, *He sells books*) and intransitive (with no object: *He reads*, *This book sells*), we find that for some verbs the intransitive cases always have the same subject as the transitive, and for other verbs the subject is always one of the objects of the transitive: thus one hardly says *The oyster reads* as one hardly says *The oyster reads poetry*, and there is hardly *The universe sells* as there is hardly *He sells the universe*. We can say that the intransitive cases of these verbs are not independent sentences, but are simply the transitive sentences plus a change: either zeroing the indefinite object, so that *He reads things* becomes *He reads*, or else zeroing the indefinite subject and replacing it by the object, so that *One sells such books easily* becomes *Such books sell easily*. In such ways many distinct sentence structures are analyzed

as being simpler known structures plus stated changes. The changes were called (partial) 'transformations', because they were mappings within the set of sentences, from simple sentences to changed ones, and from pairs of sentences to single long ones.

That these transformations were not merely a simpler way of describing the structure of sentences, but also a real property of them, is seen in the fact that word-sequences which have two distinct meanings, not due to different meanings of their words, can be explained as degeneracies in the transformations; different changes on different base sentences. Thus *Robert Frost reads smoothly* is obtainable, in one meaning, from *Robert Frost reads things smoothly*, and in the other meaning from *One reads Robert Frost smoothly*.

These transformations can be discovered for each language, as being the differences in form between two sets of sentences, roughly when the inequalities of likelihood of word-combination in one set are preserved in the other (as when we compare the likelihoods of words after *book which*, above, with the likelihoods of words after *book*). When we consider the whole set of transformations, we find as will be seen below, that they can suffice to derive the sentences of the language from a subset of short sentences. However, the variety and number of the transformations is too great for them to be fundamental elements of language structure, and indeed it has proved possible to define a very few elementary changes in sentence form, each taking place in a priori statable conditions, such that every transformation is an ordering of one or more of these changes.

These elementary changes are of few physical types: mainly, reduction of a word to zero (e.g. *I* in *I turned and left* from *I turned and I left*), reduction of a word to a pronoun (e.g. the second *the book* to *which*, above), reduction of a word to an affix (to take a very simple case: *-hood* in *child-hood* from an earlier free word *had* 'situation'). They are defined as taking place in the word *A* last entering a sentence (in the sense given below) or in the words *B* entering last before that, if the amount of information that *A* brings to the sentence over and above *B* is exceptionally small. Since the set of such *A, B* word-pairs is finite, though large, the individual reductions, which are a subset of this set, can be listed for a given language. In contrast, the set of transformations, taken as differences between subsets of sentences is not statable a priori, and possibly is not formulatable in a finitary manner. Note that the set of sentences is unbounded since there is no longest sentence. Very many of the elementary changes are optional; that is, the unchanged 'source' sentence in sayable as well as the changed

one: *I found a book; I had lost the book*, as well as *I found a book which I had lost*. We can make virtually all the remaining changes optional if we accept their 'source' sentences as grammatically possible (marked †) though not actually said, e.g. if we take *His early childhood was unhappy* from *His early situation of being a child was unhappy*, reduced to the compound-noun form †*His early child-situation was unhappy*, reduced to a suffix form *His early childhood was unhappy*.

The grammatically-possible sentences satisfy the rules of syntactic structure of all the other sentences, but are not said, either because of special difficulties with particular words (e.g. the words' having dropped out of use in free position, as with *had*, *above*, or their inability to carry particular suffixes), or because of stylistic preferences for the reduced forms.

Some of the unreduced sentences, i.e. those which are not the product of any reduction, are reconstructed from reduced sentences, as being their grammatically-possible but unsaid sources. Most are simply the sentences of the language before the optional reductions have taken place. Together, the unreduced sentences form a base for the set of sentences of the language, since all the other sentences of the language are formed from the unreduced ones by the regular application of the stated reductions. The sentences that are actually said, both unreduced and reduced, do not by themselves form a well-defined set: many are marginal (e.g. *The baby gave a crawl*), some are dubious (e.g. *I like that she should be on time*), others are said by one person but not by another. But when we include the reconstructed ones, then the unreduced (base) sentences, both those that are said and those that are reconstructed, form a well-defined set consisting of all the sentences that satisfy a certain structure, stated below. The reductions, which create out of these all the remaining sentences, take place over stated domains of the words entering a sentence. Some of these domains can be extended by the speakers, or have other imprecisions, and it is this that makes the remaining, reduction-bearing, sentences a not well-defined set.

This base set of sentences has many important properties. When we include the reconstructed sentences (marked †) in the base, then for each reduced sentence in the language there exists a base sentence which contains no reductions. Since the reductions can be seen to make no change in the information of the sentence, as in the examples above, the information they carry is carried also by their source sentence, so that all the information carried by the language is carried by the base set of sentences.

As we analyze a longer sentence by discovering what reductions they contain where, we in many cases decompose the longer sentence into shorter ones. Thus, recognizing that the *which* above is a reduction of *book* involves admitting two short sentences as the source of *The book which was lost cost \$ 5*. If we now consider the longer sentences that remain in the base, e.g. *That John writes music is probable, Mary said that John writes music*, we find that they contain a shorter sentence as proper part (*John writes music*) together with residues which are not themselves whole sentences (*is probable, Mary said*). One can see, however, that the relation of the residual words to each other or to the contained sentence is much the same as the relation among the words of the contained sentence. We therefore try to formulate that relation.

For stated sets X , Y of words, we define a relation $X > Y$ among the words of each sentence, which holds if the necessary (but not sufficient) condition for the presence of X in a sentence is the presence in it of some word of the set Y of which Y is a member. We say that X depends upon Y , or that X is later than Y in entering into the composition of the sentence. For example, certain words A , e.g. *probable, possible, continue*, occur only in sentences in which there occurs a word from a certain set B which includes e.g. *fall, write*, and also includes the words of A itself. Thus we have *That John writes music is probable, John's writing music continues*, *That John's writing music continues is probable*, but not **John is probable*. This dependence of A -words on B -words may be hard to determine in long sentences, where there may be many words of A and of B present; but it is obvious in the short sentences of the base set, and can then be recognized in all other sentences because the other sentences are decomposable into segments containing the same word-relations as do the short sentences. In contrast to the A -words, words which are in B but not in A can occur also in sentences which do not contain words of A : *John writes music, John fell*. Thus, over the whole language, $A > B$, but $B \not> A$. Another example is words A' such as *entail, because*, whose dependence is on a pair of B -words (rather than a single B); *John's writing music entails his leaving college, John's falling was because of his rushing about, That John's writing music continues entails his leaving college, That John left college is because of his writing music entailing his getting a job*, but not **John entails a job, *Music is because of college*. Here we have $A' > B, B$ (where B includes also the words of A'), but $B \not> A'$. In the base sentences, where these dependences are demonstrable, the dependent word

comes after the first of the ordered words upon whose presence it is dependent: *probable* after the sentence containing *writes*, *entails* between that sentence and the sentence containing *leave*. (Above, the *-ing* and *that* are markers indicating that there is present some word—such as *probable*, *entail*—which is dependent on the *-ing*-bearing or *that*-bearing word.)

The situation of this dependence is less clear in the minimal base sentences, i.e. those that do not contain any shorter base sentence as a proper part: e.g. *John writes music*. Here the dependence is mutual: no word seems to be more dependent than the other. Nevertheless, there is a difference among them, for the second word is similar in morphology and position to the dependent words above. Hence it is convenient to consider the second words of the minimal base sentences, such as *writes*, *falls*, *leaves*, *loses*, *finds*, to be the ones that are dependent upon the presence of their neighbors in the minimal sentences, such as *John*, *music*, *job*, *college*, *book*.

We now consider the structure of the base sentences with respect to this dependence. If we characterize words only by a partially-ordered dependence (and not a mutual dependence), then there must exist some words whose presence in a sentence does not depend on anything, for otherwise no other words—those that depend upon the presence of something else—could be present. By the same token, every base sentence must contain at least one such word. We call these words ‘primitive arguments’, *N*: *John*, *music*, *book*, etc. (But not all nouns in a language are such.) Then there must be some words whose dependence is only on primitive arguments for no other kind of word could enter a sentence which contains only primitive arguments. In English we find certain words whose presence depends upon the presence of one primitive argument, e.g. *fall*, *sleep*, *cough*, as in *John sleeps*, etc. Using O_Z to indicate words that depend on *Z*, we indicate these last by O_n . Other words (O_{nn}) depend upon two ordered primitive arguments, e.g. *write*, *lose*, *find*, *get* in *John writes music*, etc. A few (O_{nnn} , etc.) depend upon three or more—e.g. *put* in *John puts the book on the table*. The words which depend on something are called ‘operators’, and the words on which they depend, in a given sentence, are called their ‘arguments’ in that sentence. The symbol for an operator carries subscripts indicating its arguments. The words whose arguments are only primitive ones—the operators considered above—are called ‘elementary operators’. For every *m*-argument elementary

operator in a base sentence there must be present m primitive arguments which are free for that operator, i.e. which have not been counted as arguments of any other operator in the sentence.

In addition there are certain words (non-elementary operators) whose presence in a sentence depends upon the presence of one or more operators. These include the O_0 -words, such as *probable*, *continue* in *That the book fell is probable*, *John's writing music continued*, *John's sleeping continued*; the argument of the O_0 -word is an operator, which has its own arguments with it. The non-elementary operators include also the O_{n0} -words such as *entail*, *because*, which have two operators as their arguments. English also has words which depend on a pair of arguments, one an operator and the other a primitive argument, in one order or another: O_{n0} -words such as *know*, *hope*, *say* in *John knows that the book fell*, etc., and O_{0n} -words such as *astonish* in *The book's falling astonished John*. When an operator becomes an argument of a further operator, it receives *that* or *-ing* as indicator of its changed status. A fact which is of great importance to language structure is that the words which are dependent on operators make no distinction as to what is the argument-class of the operator which has become their argument. For example, *continue* does not ask whether its argument—*write*, *sleep*, *continue*, etc.—is an O_n or an O_0 or an O_{n0} , etc.; i.e. it does not depend on the argument of its argument. Thus the condition for the presence of *continue* can be satisfied not only by *write* but also by any other operator of whatever kind: something's continuing can continue, something's entailing something can continue, someone's knowing something can continue. This is one of the facts that make all properties of grammar involve no more than the relations between an operator and its arguments. And, as will be seen below, it contributes much to the mathematical character of the structure.

The base sentences can be formulated in such a way that they involve few or no word-subsets, other than N and O . If we ask what sentences are possible rather than which ones are actually said, we not only admit such cumbersome sentences as *His early situation of being a child* but also the sentences with unlikely combinations of words within the normal grammatical constructions, e.g. *He took a crawl*, *The astute ceiling thinks that we are late*. In the grammatical statement there is no point at which one could draw a line between the set of likely combinations (e.g. *He took a walk*) and the set of unlikely ones, nor are the likelihoods unchanging. For the base sentences, if we say that a certain set of words are, say, O_{n0} , i.e. that each requires the presence of a primitive argument N and an

operator O as its ordered arguments, then each of them can occur with any N and any O , as in the example of *Motors think* above; and indeed that sentence could occur in science fiction or in a joke, without being ungrammatical. Within each set, e.g. O_{no} , each operator has a partial ordering for its likelihood of occurring with each word in its (N or O) argument domains. This likelihood is imprecise, but it is preserved under all further events in the composition of the sentence, whether reductions or the entry of further operators. Indeed, if two sentence-forms show the same inequalities of likelihood for the arguments of an operator, we assume that one is the result of reductions (transformations) in the other. The restrictions and exceptions that are so familiar in grammar can be stated as limitations not on word-entry but on the domain of reductions: which words get reduced under what conditions.

The whole grammar is thus stated in terms of the operator-argument (or dependence) relation. Every occurrence of an operator on the sentence thus produced produces a further sentence in which the first one is an argument, and a proper part. As to reductions, in almost all cases, the word that gets reduced is one which contributes to the sentence little or no information, given its position over its arguments or under its operator in the sentence as so far constructed. The reductions are made upon entry; that is, (a) on a word as it enters the sentence, or (b) on one of the last entering words as an operator enters upon them, or (c) as soon as a further operator empties the informational contribution of the given word. As an example of (a): in *I request you: wash yourself!* the operator *request* with its first two arguments *I*, *you* can be zeroed, leaving the third argument *Wash yourself!*; in this rare kind of reduction, the informational grounds are the performative status of *I request you*, namely that saying *I request you* of an imperative is the same as making (saying) that imperative sentence. As an example of (b): the indefinite nouns (or so called ‘pronouns’) *something*, *things*, etc., carry little information; hence, in most cases, when they are the second argument of an operator they are zeroable, as in reducing *John reads things* to *John reads*. As an example of (c): *Boys take these jobs because of needing the money* is said only if it is the same boys that need the money; it is therefore reduced not from *Boys take these jobs because of boys’ needing the money* but from something like *Boys take these jobs because of boys’ needing the money*, where the first argument of the second argument is the same as the first of the first (i.e. where the second-mentioned boys are the same as the first). Thus it is only after the metalinguistic last sentence is added that the second

boys is zeroable. That reductions are made as soon as the conditions for them are satisfied, and are not otherwise delayed, is seen in many sentence derivations, and explains for example why pronouns are late changes in a sentence, since they depend upon that sentence being joined to another sentence and are therefore formed after the internal changes of each component sentence have been made.

In respect to meaning, the operators are predication. That is, the word whose presence depends upon certain other words says something about those other words. Those words and affixes of English which are not obviously operators or their arguments—e.g. *the*—turn out to be derivable by reduction from particular operators and arguments. All relations other than the predicational operator-argument relation—e.g. the modifier relation—can be obtained via particular reductions from the operator-argument relation. The meaning of a sentence, or rather the information carried by it, is given directly by the meaning of each of its operator-argument portions, i.e. by its elementary operators as predication on their arguments, and then by each successive further operator. Then the meaning of a sentence is not something else again, to be considered after the syntax is determined, but correlates in a regular way with the syntax of the sentence.

And now, as to the mathematical possibilities. It is possible to apply mathematics, in the usual ways, to the study of language phenomena. One can describe stochastic processes for determining word and sentence boundaries, and certain algebraic structures for sentence composition. As generally in applied mathematics, these investigations accept certain objects which are determined within a science of the real world—linguistics; they then describe the combinations or changes of these objects. However, the analysis given above makes possible something else, a mathematical characterization of language. The way to this is prepared by the elimination of restrictions and exceptions from the occurrence-dependence of words, moving these to the domains of reductions on the words. This makes it possible to consider mappings between sets of linguistic objects—at least in the unreduced sentences—without having to formulate special provisions for exceptions and the like. A further step here is the fact that the only arguments which characterize the various sets of operators are *N*, the primitive arguments, and *O*, the set of all operators. The importance of this is that these arguments are themselves defined purely by their occurrence-dependence. *N* is the set of words whose presence in a sentence is defined as not depending on anything. Within the vocabulary of the

unreduced sentences, O is the complement set, of words whose presence depends on the presence of something else— N or O in some combination. Thus, words are characterized in sets either as having null dependence (N), or as depending on one out of a few combinations of words which are in turn identified only as having null or non-null dependence. Since the word-sets are not otherwise defined, they are characterized only by their relation to words which are characterized in respect to this same relation. Within each word-set, the individual words can be identified syntactically by their inequalities of likelihoods of occurrence in respect to the individual words in their operator-set, whose words in turn can be identified by their inequalities of likelihoods of occurrence in respect to the individual words in their argument sets. We are thus dealing with sets of arbitrary objects, defined only by their participation in a relation in respect to each other—a mathematical object.

Language is a particular realization of this mathematical object with its occurrence-dependence relation, a particular interpretation of the abstract system. But any other physical system in which the combination of parts was based solely on such an occurrence-dependence relation would be language-like. And if the occurrence-relations of the new physical objects are identical in detail with those of language, we obtain a set of sequences isomorphic to the set of sentences, as indeed we have in writing vis-a-vis speech.

The structure of sentences and the relations among them can be described as certain simple algebraic structures. These are chiefly partial orderings, monoids of non-elementary operators (but not of the binary O_{00} , which are mostly non-associative in respect to meaning), and equivalence relations which provide partitions of the set of sentences. These structures are important, because every relation in them has an interpretation which is an essential part of the meanings of sentence structures; and those meanings in a sentence which are directly connected with the grammar of the sentence are interpretations of the stated relations in these algebraic structures.

The sentences of the base set are a partial order (a particular kind of semi-lattice) of arbitrary objects. The objects (in actual languages, words) have the additional property of being classifiable according to whether they occur, in the operator-argument semi-lattices (i.e. in sentences), as (1) l.u.b. only of their own occurrence in the partial order (N , elementary arguments), or (2) l.u.b. only of the latter and themselves ($O_{n...n}$, elementary operators), or (3) l.u.b. also of objects which are themselves the l.u.b. of

objects other than themselves, as well as possibly of N (these are $O_{...0...}$, non-elementary operators). That is to say, these three types of l.u.b. positions are filled in general by different objects. Furthermore, within each of the latter two position-types there are sub-types according to the number and order of N and O in the immediately lower position in the partial order: in type 2, according to how many N are immediately below it; in type 3, according to what sequence of N and O is immediately below it (O representing any object in type 2 or 3). These sub-types of l.u.b.-status are generally filled by different objects.

The set of all unary non-elementary operators, i.e. those whose argument-dependence includes precisely one operator, generates a free monoid, with successive application (i.e. next later entry) as operation, and the null operator as identity. In this, the monoid-words are products of operators $O_1 O_2 \dots O_n$ (where $O_i O_{i+1}$ means that O_{i+1} is the operator on O_i in a sentential partial order). A product of two monoid-words is itself a monoid-word; the multiplication is associative. Each monoid-word represents the succession of operators on an elementary sentence or on a binary operator on two sentences. This structure has not so far been found to be of any great importance in dealing with language. However, it illustrates how using the partially-ordered dependence-relation, instead of the overt word-sequence, makes it possible to find various mathematical structures in language. In contrast, word-concatenation in sentences is non-associative and ambiguous: *The yellow and green cards* can be derived both from *The cards which are yellow and green* and *The cards which are yellow and the cards which are green*. The entry of operators, which together with reductions describes the same sentences as concatenation would, is associative and non-ambiguous. Mappings and operations on sets of sentences can therefore be more conveniently carried out on the entries, and in particular on the operators, in the sentences than on the word-sequence of the sentences.

The binary non-elementary operators, i.e. those whose arguments include two operators, form a set of binary compositions on the set of sentences. In the base set, each binary non-elementary operator can act on every pair of sentences, although its likelihood of occurrence is lower on sentence-pairs which do not contain in their base form a word in common. In the whole set of sentences, certain statable pairs (e.g. an assertion and a question) will not appear under certain binary operators: **I am late because will you go?* In these cases we can say that the product of the two sentences (*I am late*, *Will you go?*) under the binary operator (*because*) is included

in the null sentence. Products of these binaries are in general not associative.

The reductions in a sentence act as a partially ordered set on particular operator-argument pairs, those which have the likelihood (or low information) properties required for the given reduction. Some of these reductions (if there are more than one) can be viewed as taking place simultaneously on the given operator-argument pair; others are such that one reduction operates on the resultant of another on the same operator-argument pair. Some of the large grammatical transformations such as the interrogative form and the passive are not single reductions but successions of reductions on a single operator-argument pair in a sentence or on successive operators in a sentence.

The most important algebraic structures in the set of sentences S are those which arise from equivalence relations in S in respect to the particular operator-argument semi-lattice in each sentence, and in respect to the highest operator (the upper bound of all words in the semi-lattice) or to the reductions on words in the semi-lattice. These equivalence relations identify the informational sublanguage (the base) and the grammatical transformations, as will be seen below.

We note first that the resultant of every operator is a sentence. Every unary non-elementary operator acts on a sentence (and possibly some N) to make a further sentence; and every binary non-elementary operator acts on two sentences (and possibly some N) to make a sentence. Every reduction acts on a sentence to make a (changed) sentence. All of these, in acting on a sentence, preserve the inequalities of operator-argument likelihoods in the operand sentences. The unary non-elementary operators are a set of transformations on the set of sentences: each maps the whole set of sentences S into itself (specifically, onto a subset of sentences which have that non-elementary operator as their latest entry); and the binaries map $S \times S$ into S . The reductions are a set of partial transformations on S , each mapping a subset of S (sentences containing a particular low-information operator-argument pair) onto another subset of S (sentences containing the reduction on a member of that pair).

The preservation of inequalities of likelihood under transformations, i.e. under the non-elementary operators and under the reductions, is of great importance. Without it, there would be no semantic connection between a sentence and its occurrence under further operators or reductions. The operators preserve the likelihood-inequalities and the meanings in their operand sentences, although with a reasonable number

of specified exceptions. The non-elementary operators also add their own meanings and likelihoods in respect to their argument, so that the inequalities among the resultant sentences (with their new higher operator) need not be the same as among the corresponding operand sentences. As to the reductions, they preserve with only few if any exceptions the inequalities of likelihood and the meaning of their operand sentences and add no objective information to it; they are paraphrastic. Here too the reduction may raise (or lower) the likelihood of occurrence in the resultant sentences, but for the most part equally on all its operand sentences.

We now consider the set S , where each word-sequence which is grammatically ambiguous in n different ways is considered to be a case of n different sentences. S is a semi-group under the binary operator *and*: for any two sentences A, B we have A *and* B as a new sentence C . (This, after adjustments are made for the stable A, B pairs which do not take *and*.)

We present now a structure which isolates the minimal subset of the set of sentences as a residue of the non-elementary operators and reductional transformations. It has little importance when the great bulk of transformations are products of such simple reductions as have been established for English, and when the minimal sentences can be characterized, as has been done here, as the resultants of elementary operators on primitive arguments. However, in a language in which we do not have so clear a picture of the structure of the set of transformations or of the set of base sentences, such as a way of identifying the minimal sentences is useful. To obtain this structure, we take an equivalence relation in S , whereby two sentences are in the same equivalence class if they contain the traces of (i.e. exhibit the presence of) the same monoid-word of unary operators and the same partial orderings of particular reductions. There is a corresponding binary composition in the set of equivalence classes E , with E_A *and* $E_B = E_{A \text{ and } B}$ (where E_X is the equivalence class to which the sentence X belongs). In the natural mapping of S onto its quotient set E , the kernel of the mapping, i.e. the sentences which are mapped onto the identity of E , includes elementary sentences and also the resultants of the binaries. In each of these resultants, the two operand sentences are then assigned to equivalence classes in the same manner as the original sentences. When no resultants of binaries are left in the kernel of the natural mapping, this sub-kernel contains only minimal sentences.

A different and much more important structure is obtained if in the set S we take an equivalence relation by which two sentences are in the

same equivalence class if they have the same ordered word-entries (i.e. the same operator-argument semi-lattice). Since almost all reductions are optional, each equivalence class contains (with possibly certain adjustments) precisely one reduction-less sentence. The set of these is the base set, from which the other sentences are derived by reductions. The base set is closed under the word-entry operation: any word sequence satisfying this form is such a sentence. Hence we may call this set a 'sublanguage'. Since the domains of successive reductions are monotonically decreasing, the base set of sentences, one from each equivalence class above, is the most unrestricted in respect to word domain.

There are many properties of language that can be derived from this analysis, or are clarified by it. One is that since the base set, which suffices for all the information carried by language, has virtually no restrictions or exceptions and lacks all the special constructions of grammar such as conjunctions, tenses, etc., it follows that, contrary to common views, all these are not essential for expressing the information carried in language, nor are they essential for language. Another is that since the whole structure of language is seen to be predicational, it is clear that language developed as a tool for communicating information rather than purely as a form of expression. Yet another is that since the whole of language arises in explicable ways from so simple a relation as the dependence of word-occurrences, there is no need to assume any inexplicable structuralism underlying language.

FORMAL CAPACITY OF MONTAGUE GRAMMARS

CARL H. HEIDRICH

University of Bonn, IKP, Bonn, F.R.G.

1. Introduction

Since MONTAGUE's now classical papers *English as a formal language*, (EFL—1970), *Universal grammar*, (UG—1970), and *The proper treatment of quantification in ordinary English*, (PTQ—1973) have been published, the literature on Montague's theory, a treatment of semiotic as it is carried out for instance in UG, falls roughly into the following classes or intersections of these classes:

- Extending the fragments of English language
(Bennett, Partee, Thomason, etc.)
- Expositions of the main ideas of Montague's theory especially for linguists
(Hamblin, Lewis, Partee, Thomason, etc.)
- Special logico-linguistic problems
(Bennett, Hintikka, Kamp, Kaplan, Karttunen, Partee, Thomason, etc.)
- General or special discussions of pragmatical theories
(Cocchiarella, Gabbay, Gallin, Kamp, Kaplan, Kripke, Lewis, Scott, etc.)
- Treatments of transformational-linguistic problems within Montague's theory as well as theory comparison
(Cooper and Parsons, Dowty, Partee, Rodman, etc.)
- The class of the images of the aforementioned classes formulated in or for other languages than English.
- The critical or polemical positions
(Kasher, Martin, Potts, etc.)

Linguistic research using Montague's theory, for instance by use of a Montague grammar as a tool for descriptions of syntactical properties

of a fragment of a natural language, is rapidly growing and is getting very complicated as can be seen from the recent literature, compare e.g. M. Bennett's article on demonstratives (*Synthese* 39, 1978). Since Montague's theory is formulated (in UG) in a strict mathematical manner, it suggests itself to reflect upon the premises, assumptions and basic properties of the algebraic treatment. This may help us to construct new fragments of natural languages within Montague's theory, which must not necessarily be extensions of fragments already known.

2. Montague theory-appeal

Unfortunately, some of Montague's statements concerning the relation between formal and natural languages, or his evaluation of the "developments emanating from the MIT" have irritated and bewildered people, especially linguists and philosophers of language. On the one hand, this incident should be seen more in a social background. On the other hand, Montague's statements have an eminent argumentative force, even if they are evaluated only rhetorically. But the reasons for his statements appear as methodological conclusions drawn on the basis of the algebraic constructions of his theories for languages.

I assume that Montague's statements have been misunderstood. He states in EFL as well as in UG that there is no *theoretical* difference between natural and formal (or artificial) languages. Here "theoretical" means: both types of languages can be or should be described by (or: be formulated in) *one type*—not different types—of algebraic structures, e.g. as disambiguated languages defined in UG. Montague's statements do not imply or do not presuppose that there is no empirical or other sort of difference between natural and formal languages. Montague's statement does imply that the natural language objects occurring in Montague grammars are formal objects.

A similar explanation can be given to the concept "universal grammar". It is used by Montague in a somewhat metaphorical sense. It refers to the traditional concept of linguistics. But, on the other hand, it refers to mathematical theory and is a weighty program: Algebraic structures and algebraic constructions shall be used in the study of natural as well as formal languages. Thereby the study of languages shall become part of *universal algebra*, as a mathematical discipline, as large parts of model theory for logical languages are already.

The justification for his statements and his program can be found in

an important passage of his *On the nature of certain philosophical entities*. He expressed his conviction, he called it "my dogmatism", already in 1967.

The two decisive statements are, first,

"...that philosophy, at this stage in history, has as its proper *theoretical* framework set theory with individuals and the possible addition of empirical predicates." (my emphasis, CHH.).

and second,

"Philosophy is always capable of enlarging itself; that is, by meta-mathematical or model-theoretical means—means available within set theory—one can 'justify' a language or a theory that transcends set theory, and then proceed to transact a new branch of philosophy within the new language." (MONTAGUE, R., *Formal Philosophy*, ed. by R. THOMASON, 1974, p. 154, 155, respectively).

The concepts *Montague-theory* and *Montague-grammar* occurred, to the best of my knowledge, in lectures and publications of Partee, compare e.g. *Montague Grammar and Transformational Grammar*, Linguistic Inquiry VI, 1975, and can perhaps be traced back to 1972. Though, with respect to UG, an immediate comprehension of both concepts can be assumed, no explicit definitions for both concepts have been given. The definitions depend on the concept of an algebra:

- 0** The algebra \mathcal{A} of type τ is an ordered pair $\mathcal{A} = \langle A, \langle f_i \rangle_{i \in I} \rangle$, A a set, $\langle f_i \rangle_{i \in I}$ a family of operations of type $\tau = \langle \alpha_i \rangle_{i \in I}$ for any set I .
- 1a** A *Montague-grammar* (relative to a natural or artificial language N) is a language $L = \langle \mathcal{A}, R \rangle$ as defined in UG, where $\mathcal{A} = \langle A, \langle F_\gamma \rangle_{\gamma \in \Gamma}, \langle X_\delta \rangle_{\delta \in \Delta}, S, \delta_0 \rangle$ is a disambiguated language (relative to a natural or artificial language N).

As one can see, the category structure X_δ of a Montague-grammar has to appear explicitly as a family of sets indexed by category indices besides the indexed family of the operation structure and the rules S . A few notational differences occur in EFL, UG, and PTQ. In EFL the category concept is applied to syntactical as well as to semantical objects. With respect to logical languages the operations of the algebra are applied in UG, PTQ to objects of different *types*, which are in fact semantically determined. One has to keep in mind two readings of the word "type": the one, as in **0**, to discriminate the operations occurring in an algebra

according to their place numbers, and the other, to indicate the (semantical) status of an object under an operation.

The definition of the second concept has to take into account the different methods of interpretation of natural or artificial languages applied in EFL and in UG, PTQ. Considering the method of direct interpretation of EFL, the correlation between syntactical relations of Montague-grammars and (functions on) possible denotations is based on a model (or relative to a model). In the case of the method of translation used in UG and PTQ the syntactical relations of Montague-grammars (of a natural language) and syntactical relations of Montague-grammars of formal intensional languages are correlated relative to a translation base. The intensional language is given a direct interpretation. The intensional language, its direct interpretation, and the translation base induce an indirect interpretation of the Montague-grammar (of the natural language).

The following definition of the concept Montague-theory uses the terminology of UG. I shall leave out the definitional clauses of the definiens-concepts in the definition. They have to be supplemented according to the definitional clauses for those definiens-concepts as they are given in UG.

1b (EFL) A *Montague-theory* (for a language L) is a system

$\mathcal{L} = \langle L, T, \sigma, D, \langle \mathcal{B}, r \rangle \rangle$, such that

L is a Montague-grammar,

T is a set of (semantical) types,

σ is a function (type assignment) from A into T , A is the set of category indices of A of L ,

$D = D_{T,E,P}$ is the class of possible denotations relative to the set of types T , a set of entities E , and a set of points of references P ,

$\langle \mathcal{B}, r \rangle$ is a model for L such that

r is a point of reference,

$\mathcal{B} = \langle B, \langle G_\gamma \rangle_{\gamma \in \Gamma}, f \rangle$ is a Fregean interpretation for L , such that
 $B \subseteq D$.

1b (UG) A *Montague-theory* (for a language L) is a system

$\mathcal{L} = \langle L, L', \mathcal{T}, T, \sigma, D, \langle \mathcal{B}, r \rangle \rangle$ such that

L, L' are Montague-grammars,

$\mathcal{T} = \langle g, \langle H_\gamma \rangle_{\gamma \in \Gamma}, j \rangle$ is a translation base from L into L' ,

T is a set of (semantical) types,

σ is a function (type assignment) from A' into T , A' is the set of category indices of A' of L' ,

$D = D_{T,E,P}$... (as in the foregoing definition),
 $\langle \mathcal{B}, r \rangle$ is a model for L' ... (as in the foregoing definition),
 $\mathcal{L} = \langle L', T, \sigma, D, \langle \mathcal{B}, r \rangle \rangle$ is a Montague-theory (for the language L'). (e.g. relative to 1b (EFL)).

From Montague's publications one cannot conclude whether or not he would like to identify the formal objects of a Montague-grammar with those natural language objects as they are described by informal grammars. Standardized semiotic properties are normally used to establish the apparatus of the categories of (formal) Montague-grammars as they are used at the moment to describe fragments of natural languages. But usually the fundamental standardizations have not been stated explicitly. Of course, one can find them in the fragments. In EFL, UG, PTQ it is presupposed that the (semiotic) standardizations are lexematically and morphologically represented. A standardization is a decision whether an element is e.g. contained in a specific set of a certain base category, for instance in the category of the ad-one-verb phrases, or not. It may also be a rule that introduces a morphological or lexical object not contained in any category, e.g. 'such that' is introduced by a rule, S 16, in EFL.

At first, from the formal point of view, syntactical categories C and semantical types T do stand apart. (I shall use both concepts from now on in the sense of PTQ, i.e. $C = \text{CAT}$, $T = \text{Type}$, the sets of categories and types respectively. C , T serve as index sets for sets of expressions.) If a syntactic-semantic parallelism exists, what is the mathematical way to manipulate it?

Here I shall not discuss the philosophical questions concerning the existence and necessity of a syntactic-semantic parallelism. The answer to this question depends primarily on the structure of the syntactical categories C and the semantical types T . We have to map, e.g., the syntactical categories onto the semantical types such that syntactical rules correspond to semantical functions. As we shall see in the next section, all (functional) semantical types can be expressed as polynomial operations on semantical types. Then, the polynomial operations on semantical types induce polynomial operations on (functions on) possible denotations, i.e. on elements of a universe of entities of respective types. The correspondence between syntactical rules and semantical functions is explicitly expressible if semantical functions are represented by polynomial operations on possible denotations, such that the types of possible denotations correspond to the categories occurring in the syntactical rules.

The manipulation of the polynomial operations may be a very intricate subject if a syntactically and semantically rich language is given.

This construction secures the existence of strong relations, homomorphisms, between the algebra of a language and the algebra of the functions on possible denotations, if the precaution is taken that the two algebras are similar, that is, the operations are of the same type. (Remember, "type" is here used in the first of the above mentioned readings.) One finds this precaution expressed in UG in the definition of "meaning assignment". It is also contained in the above definitions of the concept Montague-theory if they are written out completely. Altogether, the method of Montague's construction turns out as an extension of standard constructions of universal algebra.

One consideration is important here too: if confronted with the question whether syntactical categories have priority over semantical types or vice versa, one has to answer it from the point of view of the algebraic construction: one can either establish a homomorphism from language to denotations or from denotations to language. In both cases one needs polynomial operations with respect to types or to categories to establish a syntactic-semantic parallelism.

3. Semantical types

The remarks of Section 2 should point out the significance of the semantical types in constructions of Montague-theories. This section contains some observations formulated in a somewhat informal manner having its origin in properties of semantical types. Some consequences for constructions of Montague-theories for fragments of natural languages will be illustrated.¹

Relations-in-intensions or *I-predicates* are roughly functions from possible worlds into sets of entities, (more exactly, e.g. functions into the set of all subsets of the Cartesian product of sets of entities). Possible worlds and sets of entities, *I*-predicates as well as extensional relations between entities are the building-stones of Montague's ontology. The *I*-predicates which may or may not hold for different kinds of entities are standardized

¹ A technical representation had to investigate relations between Peano algebras, for an algebra of semantical types as well as an algebra of a disambiguated language (relative to this algebra of semantical types), is a Peano algebra: both are algebras as defined in the present context.

according to types. Hence an intensional language will be a type-theoretical one.

Assume for the following considerations that there exist possible worlds and universes of entities, one may then also assume the existence of sets. Entities are classified by Montague according to types and possibly by possible worlds. If T_s is a type structure over a set of types T , then the classification ζ is a mapping from T into E , the union of the universes of entities, such that 2 holds for some or perhaps for any $t \in T$ and $U \in E$.

$$\begin{aligned} 2 \quad & \zeta: T \longrightarrow E \\ & t \longmapsto U_t \\ \text{or: } & \zeta(t) := U_t. \end{aligned}$$

There is a problem with the phrase "for some or perhaps for any". I shall return to this later.

The construction of the type structures T_s is an important subject. A simple construction would be to identify T with N , the set of natural numbers. This is normally not very useful, because, as Ajdukiewicz has suggested, types should be constructed as functionally related objects. The construction underlying Montague's theory can be done as follows: choose three different objects, say 0, 1, 2, and let $e := 0$, $t := 1$, $s := 2$, let $M = \{0, 1\}$, i.e. $s \notin M$.

3 T_s , the algebra of types, is

$T_s = \langle T, \langle f_0^2, f_1^1 \rangle \rangle$ containing one 2- and one 1-placed (partial or full) operations f_0^2, f_1^1 and is the free algebra generated by M , that is, the following four conditions are satisfied:

- i. $M \subset T$,
- ii. $\bar{M} = T$, or, T is the smallest set containing M and is closed under operations f_0^2, f_1^1 ,
- iii. range $(f_i) \cap M = \emptyset$, $i = 0, 1$,
- iv. if $f_i(a) = f_j(b)$, then $i = j$ and $a = b$, $a \in \text{domain } (f_i)$, $b \in \text{domain } (f_j)$, $i, j \in \{0, 1\}$.

One gets the types of PTQ by the following definition.

$$\begin{aligned} 4 \quad & \langle a, b \rangle := f_0^2(a, b), \quad a, b \in T, \\ & \langle s, a \rangle := f_1^1(s, a), \quad a \in T. \end{aligned}$$

Since $M \subset T$, we have of course $e, t \in T$, but $s \notin T$.

T_s contains an infinite number of functional types. For instance, all of the following sequences are in T_s .

5a

| | |
|-----|--|
| t | |
| | $\langle t, t \rangle$ |
| | $\langle t, \langle t, t \rangle \rangle$ |
| | $\langle\langle t, \langle t, t \rangle \rangle, t \rangle$ |
| | $\langle t, \langle\langle t, \langle t, t \rangle \rangle, t \rangle \rangle$ |
| . | . |
| . | . |
| . | . |

b

| | |
|------------------------|---|
| $\langle s, t \rangle$ | |
| | $\langle\langle s, t \rangle, \langle s, t \rangle \rangle$ |
| | $\langle\langle s, t \rangle, \langle\langle s, t \rangle, \langle s, t \rangle \rangle \rangle$ |
| | $\langle\langle\langle s, t \rangle, \langle\langle s, t \rangle, \langle s, t \rangle \rangle \rangle, \langle s, t \rangle \rangle$ |
| . | . |
| . | . |
| . | . |

I mention a few other ones which are in fact special types of PTQ. One may form an idea of the complicated constructions of functional types which can be reached in few steps, eight steps in the example, of applications of the operations of Ts.

- 6a**
1. $\langle s, e \rangle$ individual concepts,
 2. $\langle\langle s, e \rangle, t \rangle$ sets, characteristic functions,
 3. $\langle s, \langle\langle s, e \rangle, t \rangle \rangle$ properties of individual concepts,
 4. $\langle e, t \rangle$ sets,
 5. $\langle s, \langle e, t \rangle \rangle$ properties of entities,
 6. $\langle\langle s, \langle e, t \rangle \rangle, t \rangle$ sets of properties ...
 7. $\langle\langle s, \langle\langle s, e \rangle, t \rangle \rangle, \langle\langle s, \langle e, t \rangle \rangle, t \rangle \rangle$ relations between properties of individual concepts and properties of entities,
 8. $\langle s, \langle\langle s, \langle\langle s, e \rangle, t \rangle \rangle, \langle\langle s, \langle e, t \rangle \rangle, t \rangle \rangle \rangle$ properties of relations between properties of individual concepts and properties of entities.

A representation for the respective sets of the types involved is given in **6b**, $E = O_e \cup 2_t$, O_e the set of entities of type e , 2_t any two-element set of type t , W the set of possible worlds.

6b

$$(2^{(2^{(0^W)^W} \times 2^{(2^0)^W})^W})^W.$$

An inspection of the process generating the types in the examples just mentioned leads to the application of the concept of polynomial operations, i.e. composite operations. As said in Section 2, polynomial operations play a prominent role in Montague's theory. Take first as an example line 3 of 5a as a composite operation for three arguments: the argument t and the pair of arguments (t, t) . One of the operations in the composition is of course the fundamental operation $f_0^2 := \langle , \rangle$, the other is a constant operation $c_t^n(a) = t$ for $a \in T^n$. The composition operation is defined as

$$\begin{aligned} 7 \langle c_t^1, \langle , \rangle \rangle^T(t, (t, t)) &:= \langle c_t^1(t), \langle t, t \rangle \rangle \\ &= \langle t, \langle t, t \rangle \rangle. \end{aligned}$$

This is the composition of functions with unequal argument sequences. In the general case one has composition of functions with equal argument sequences also:

A polynomial operation $f^{An}(g_k^m)$ is an m -placed operation such that $f^{An}(g_k^m| k < n)(a) := f^n(g_k^m(a)| k < n)$,

f^n an n -placed operation over A , g_k^m m -placed operations over A , for $k < n$, and a an m -placed sequence of A elements.

As one immediately realizes, a polynomial operation can be defined for any functional type.

I have mentioned the examples to show that there are very interesting subsets T' of T . With a look at logical operators one may think for instance of the types of iterated modalities. Such subsets constitute in a sense pure fragments of our ontology. The corresponding fragments of languages for those ontologies are logical languages. Thus the concept "fragment" in Montague's theory is mainly determined by a suitable ontology with respect to types.

A language based on an ontology with respect to the full structure T s is so highly complex that we may be unable to interpret it with intuitive concepts of our natural language. What Montague has shown in his three classical papers is roughly speaking that for ontologies relative to three subsets of T he can correlate the ontology with natural language concepts. Whether any ontology E relative to a subset of T can be correlated with concepts of natural languages is an open question. I shall come to this point later again.

A further important point on T s structures is the following. How can we establish postulate systems on T s and rules within T s? To illustrate this for an example from 5a, let $M' = M \cup \{tt\}$, i.e. M' contains a pair of t 's. (Alternatively one can define a third fundamental operation on M .)

All of $\langle t, tt \rangle$, $\langle tt, t \rangle$, $\langle tt, tt \rangle$ are now elements of T_s . A postulate may be

$$8 \quad \langle t, \langle t, t \rangle \rangle = \langle t, tt \rangle$$

and this expresses a sort of a Schönfinkel-reduction for functional types. Rules are operators on T_s . For instance,

$$9 \quad R_1^2(\langle t, t \rangle, t) = t,$$

$$R_2^2(\langle t, \langle t, t \rangle \rangle, t) = \langle t, t \rangle.$$

These rules remind us of the grammatical rules. The ideas expressed in 8 and 9 are not new. They can be attributed to Ajdukiewicz and Bar-Hillel and, as far as I know, it was especially Lambek who did mathematical research on T_s algebras.² The use of the postulates and rules is expressed by the attempt to know about the similarities between types and about the reductions of type functions to simpler types.

If one extends 8 and 9 according to the other types, as for instance listed in 6a, b, then these principles may be called with respect to relations between ontology and language *Montague's constraints*. This can be seen by his treatment of transitive and intransitive verbs, or certain common nouns and adjective constructions. My view is that we should not feel fixed to or happy with this ideology, which is of course compatible with Frege's functionality principle. One critical point is that some philosophers of language believe that those constraints rest on ontological laws and purely mathematical principles—may be Frege believed this too—whereas others believe that they may be justified by empirical reasons. To this argument one can reduce some of the critical remarks about Montague's theory. Another important point is: if there are certain similarities between types, does this imply similarity between rules? Here we are at the root of a lot of discrepancies. A favorite of Montague's examples is quantification. The following comments may illustrate the foregoing remarks.

If $\langle e/t \rangle$ is a new type to be correlated with common nouns, and q is a type for quantifiers, $[q, a]$ is a new type function on T for suitable $a \in T$. The following equations express some postulates and rules for the new types.

$$10 \quad R_1(\langle\langle s, \langle e, t \rangle \rangle, t \rangle, \langle s, \langle e, t \rangle \rangle) = t,$$

$$\text{i. } R(\langle e/t \rangle, q) = \langle\langle s, \langle e, t \rangle \rangle, t \rangle,$$

² See LAMBEK, J., 1958, *The mathematics of sentence structure*, American Mathematical Monthly, vol. 65, pp. 154–170.

- ii. $[q, \langle e/t \rangle] = \langle \langle s, \langle e, t \rangle \rangle, t \rangle,$
 $R_2([q, \langle e/t \rangle], \langle s, \langle e, t \rangle \rangle) = R_1(\langle \langle s, \langle e, t \rangle \rangle, t \rangle, \langle s, \langle e, t \rangle \rangle).$

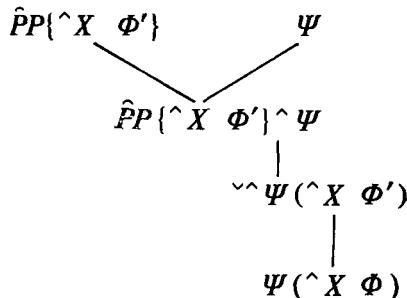
At first sight one does not see the problem of a specific Montague constraint for quantifiers and terms, because Montague expresses the similarity of 10ii by a syntactical rule equivalent to 10i in the fragment of English as well as in the intensional language and a translation rule, compare e.g. in PTQ rule S2, number (5) of the recursive definition of the meaningful expressions of the intensional language, and translation rule T2. But relative to Ts the *R* of 10i is in a strict sense no rule, it is a similarity and the correct form in 10ii expresses that quantifier terms are to be related to sets of properties of entities. A similarity rule between the rules R_1 and R_2 is then to be established on the similarity of types as it is expressed in 10ii. This similarity secures that quantifier terms and terms can be handled in a similar way. Phenomena of like kind can be shown for other rules of EFL, UG, and PTQ.

The lesson to learn from what I have said is that we must not resign ourselves to Montague's constraints in the present form. We could take more than two fundamental types for M and possibly other type functions. For practical purposes one may leave out similarities in considerations of some very small fragments of languages. The following example may illustrate a rough analysis of a sentence scheme.

11a X —such that $\Phi - \Psi$.

Let ' $X\Phi$ ' be an abbreviation for ' X such that Φ ' and let $X\Phi' = \hat{P}P\{\wedge X\Phi'\}$. With rules analogous to rules of PTQ we may have the following analysis tree for 11a.

11b



For this analysis one needs a new type, say e' , for phrases $X\Phi'$. We may later ask for what language fragments $e' = et$.

I mention a few other examples without going into the details. Decomposition of predicates, meaning postulates and transformations may be based on similarities of types and of rules. Compare e.g. Dowty's discussion of lexical decomposition and the handling of transformations in the Cooper-Parsons syntax.

This of course leads us to a fundamental question. How are *I*-predicates and possible worlds to be constructed? I am going to make some remarks on *I*-predicates.

For Montague the existence of the set *E* of universes of entities does not only result from the fact that sets exist. If sets exist and if a theory of sets postulates the existence of subsets, then one has to assume the existence of subsets of *E* for any type. Montague's position in *On the nature of certain philosophical entities* (NPE) was that we have to establish relations between ontology and certain models *U* for it.—I cannot answer the exegetical question whether he took this view all his life.—This relation is called *reduction of ontological entities to I-predicates*. Only reductions guarantee the existence of an *I*-predicate of type *a* in a model, i.e. reductions lead in a natural way to the types of *T* such that an *I*-predicate is in the model. The trouble with this method is that it was not given attention later on neither by Montague nor by others (except by myself, as far as I know)³. This fact may depend on the technicalities of NPE. A far-reaching consequence is that the intensional semantics of UG and PTQ uses *I*-predicates extensively and the type structures of the expressions of intensional logics are the outcome of the *I*-predicates which are used in interpretations of natural languages. Apparently it is accepted, especially among linguists, that there is an ontological entity for each type of *I*-predicate. This may simply be wrong or be true only in a pure mathematical sense. For instance, Cresswell assumes the existence of properties for any type in his theory of λ -categorical languages.

I shall sketch now a reduction by an informal example. The reduction is based on natural language locutions which denote or refer to entities in a natural and specific way. The example is:

- i. Bill lifts a stone at 9 a.m.

This, according to our language use, expresses a *task*.

- ii. This task is a task of Bill, 'Bill' denotes an object *x*.

³ Compare Chapter 4 of my *Intensionale Analysen von Sprechhandlungen* (Hamburg, 1977).

- iii. The expression 'to lift a stone at 9 a.m.' denotes an object R which is the task that x performs.
- iv. The expression i. is true iff x performs R , i.e. x performs a (the) lifting of a stone at 9 a.m.
- v. The task of x , x lifts a stone at 9 a.m., is the I -predicate R to x to which the entity task is reduced.

In this formulation the task R is an I -predicate of persons. It needs to be shown that the reductions can be treated formally and that procedures for handling them can be elaborated.⁴ I shall only describe the analytical parts which must be used in reductions. One needs in the case of reducing a task to a 2-placed I -predicate:

1. The entity *task*.
2. A representation for it: 'lifting a stone'.
3. Statements relative to 1., 2.: 'lifting a stone is a task'.
4. Special representations: 'lifting a stone at a given moment'.
5. A notation for it: ' x lifts a stone at t ' as well as a variable notation: $R(x, t)$.
6. Suitable notations for the special representation of the entity if it occurs:
 - x performs the (a) lifting of a stone at t ,
 - x performs the task of lifting a stone at t ,
 - x performs R at t ,
 - x performs the task R at t .

May I just hint at the second fundamental problem mentioned above: It depends on the possible worlds whether reductions are to 1-, 2-, or n -placed I -predicates.

One needs an intensional language to express properties for the reduced elements. Within a language similar to the one in NPE one property may be

$$12 \quad \hat{R} \text{ Task } (R) = \lambda t R(x, t),$$

i.e. the property of being a task is a property of 2-placed I -predicates.

As one realizes, special predicate constants like 'Task' are used to express the property of being reduced to an I -predicate. I can make

⁴ See p. 104 ff., op. cit., note 3.

a point here. I have never seen an occurrence of a predicate constant like 'Task' in any fragments for natural language constructed by Montague grammarians. (The exception is again my own publication.) This is due to the fact that the connections of NPE and UG, PTQ is almost forgotten. Partee says concerning such expressions as '^ walk', i.e. something like 'to walk' or 'the property (of x) to walk'

"I have seen none so far in any derivations."⁵

Of course not, since such sentences as

† TO WALK IS AN ACTION

have not appeared in any fragment. If one tries to write a Montague grammar for that, one will soon see that reductions must be done. For the usual types and categories will not do the job. One needs other types and categories. Let me offer a classical example of linguistics as a puzzle for Montague grammarians:⁶

IT IS EASY TO PLEASE JOHN

A solution runs along the following line.

The expression

†† TO PLEASE JOHN

is a higher order sentence, something like

THE-TO-PLEASE IS JOHN'S

A crude one of course!

IT IS EASY is a modal operator on a higher order sentence. The lower order sentence is of course

JOHN IS TO PLEASE

or

**JOHN HAS THE PROPERTY TO BE PLEASED
JOHN HAS THE PROPERTY OF BEING PLEASED**

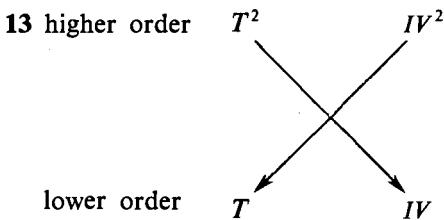
⁵ PARTEE, B., *Montague grammar and transformational grammar*, Linguistic Inquiry, vol. VI, p. 293.

⁶ See p. 82 ff, 114 f, 157 ff, n. 58 op. cit. note 3.

The modal sentence of the lower order is then

IT IS EASY (TO REACH) (FOR SOMEONE) THAT JOHN HAS
THE PROPERTY OF BEING PLEASED

If one sees the point, a principle for handling sentences \dagger , \ddagger in reductions is easy to find. Intuitively it amounts to an interchange of higher order and lower order terms and intransitive verbs.



The higher order T^2 TO PLEASE is one that can occur as an argument of a predicate constant C (like 'Task' in the other example) so that one has

TO PLEASE IS A C

In principle, the entity of a reduction and statements about it are higher order expressions and the locutions with which we indicate them are lower order sentences which reflect the properties of higher order sentences. This has immediate consequences on the type and category structures. Reductions can be formulated as those rules which regulate the interchanges of higher and lower order language objects.

It is also true that higher order sentences reflect properties of locutions of lower order. The interchange of higher order and lower order categories as in 13 must be compatible with the syntactic-semantic parallelism thesis. The difficulty in the puzzle originates from the fact that this has not been noticed.

The example as well as others clearly indicate the usefulness of reductions, i.e. special constructions of I -predicates. We may get semantical types not yet considered. It is also not clear in advance whether Montague constraints can be applied to the new types. This may be true for all cases where we have homogeneous interchanges as in the case of higher order T^2 , IV^2 with lower order IV , T , respectively. At the moment we are left with constructions of individual reductions, for there is no general procedure known to handle them.

4. Applications

In the following example I sketch how performative utterances or speech acts can be represented in a Montague-grammar. First, I present the construction and then the justification.

14 $B_{T''} = \{I : \text{YOU}, \text{JOHN} : \text{BILL}, \dots, he_i : he_{j_0}, \dots\}$ "pairs of individual terms",

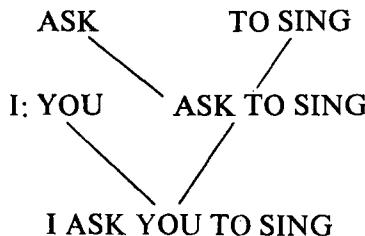
$B_{IV'} = \emptyset$ "intransitive verbs" with $IV' = t/ee$,

$B_{TV'} = \{\text{ASK}, \dots\}$ "transitive explicit performative verbs", $TV' = IV'/T'$,

$B_{T'} = \{\text{TO SING}, \dots\}$ "property terms", $T' = IV$.

The basic sets of expressions 14 for a fragment of performative utterances may be contained in a suitable fragment where the unprimed categories are as usual. The definition of C is such that $ee \in C$. It should also be noticed that use has been made of certain similarities between categories as the identities in 14 indicate. They depend upon the reductions for speech acts. The rules are such that one has among others the following analysis tree.

15



The intensional language contains among other expressions

16 $u : v, j_1 : j_2 \in \text{Con}_{\langle e, e \rangle},$

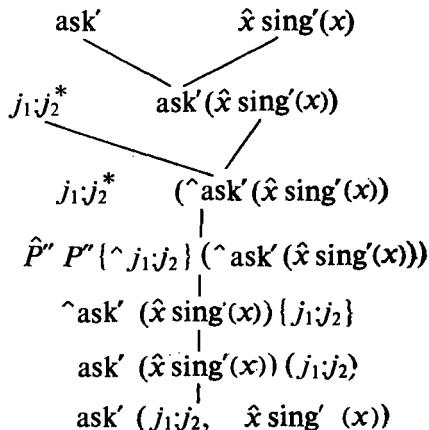
$\hat{u} : v, \hat{j}_1 : j_2 \in \text{Con}_{\langle s, \langle e, e \rangle \rangle},$

$P'' \in \text{Var}_{\langle s, \langle \langle s, \langle e, e \rangle \rangle, t \rangle \rangle} =: \text{Var}_{p''},$

$\text{ask}' \in \text{Con}_{\langle \langle s, \langle \langle s, e \rangle, t \rangle \rangle, \langle s, \langle e, e \rangle \rangle \rangle},$

$j_1 : j_2^*$ is of type $\langle p'', t \rangle$.

The rules of the intensional language are such that one gets the following analysis tree



15 can be translated into **17** if especially

18 $ee \mapsto \langle e, e \rangle$,

ASK translates into ask',

I : YOU translates into $j_1 : j^* = \hat{P}' P'' \{ \hat{j}_1 : j_2 \}$,

TO SING translates into $\hat{x}\text{sing}'(x)$.

The arguments which lead to this construction can in principle be seen in **17**. ASK TO SING should be taken as a property of pairs of persons. I explain speech acts like **15** as an act of two persons, i.e. a property P of h of being a performative utterance of a person (speaker) S addressed to a person (hearer) H . This property is different from: a property P of a performative utterance h to a speaker S of uttering h . The first property is of course not expressed in **17** but only reflected from a higher order sentence. In **17** what is expressed is only that TO ASK TO SING is one of the properties in the property set belonging to pairs of persons. The higher order sentence expresses then facts like that the property TO ASK TO SING is a speech act (containing a property) which is performed between two persons. For, as one knows, in absence of an addressee a speaker cannot in uttering a sentence perform a speech act addressed to the addressee.

The interchange of types of the expressions, i.e. higher order TV , IV , and T with lower order TV , IV , and T , involved in the reductions of speech acts leads to properties of pairs of persons, finally to the property-terms

and to the transitive explicit performative verbs. Then one is justified in establishing for our example of natural language an analysis like 17.

One may ask also how to explain act-types like THE ASKING. Concisely, the act-type THE ASKING is a relation between two generic objects, THE SINGING and THE PROPERTIES OF PAIRS OF PERSONS e.g. One may see another point here also. A reduction of actions to *I*-predicates is to be seen as a partial reduction of more complicated entities like dialogues for instance. These may be seen as sequences of actions together with situations of persons involved in common actions performed in the dialogues. So, one has to reduce dialogues to set theoretical objects which are to be taken as models of linguistic representations of dialogues.⁷

⁷ Reference can be made here to the joint work of M. Lutz-Hensel, A. Günther and myself as well as to independent research by my colleagues. Compare A. GÜNTHER, *Dialogkonstruktionen auf der Basis logischer Ableitungen* (Hamburg, 1977), C. H. HEIDRICH (ed.), *Konstituenten dialogischer Kommunikation* (Hamburg, 1977). Lutz-Hensel has done most of the research concerning linguistic analysis of performatives. ASK is only one out of four classes. Here it is a member of the performative verb class not dependent on other performative verbs with explicit addressee, a temporal present tense and a suitable semantic-syntactic classification of the complement sentences. Complement sentences are classified into 13 different classes. The results of Lutz-Hensel will be published soonly. I shall publish some material on dialogues in Montague-grammar in the proceedings of a Symposium on Theory of Argumentation, held at Groningen, 11–13, October 1978 under the title *Montague-grammars for argumentative dialogues*, to appear in: Criticism and defence, eds. E. M. BARTH & J. L. MARTENS (Amsterdam, 1980).

ON LOGICO-LINGUISTICS: STRUCTURE, TRANSFORMATION AND PARAPHRASE

R. M. MARTIN

New York, N.Y., U.S.A.

Linguists on the whole have been slow to embrace the vast riches of modern mathematical logic as the *logica utens* for their work. The situation in linguistics is rather like that in mathematics about seventy years ago. In spite of the staggering achievements of Frege, Zermelo, and Whitehead and Russell, mathematicians as a whole were not convinced of the possibility of a logical unification and systematization of their subject even up into the 1940's. Perhaps more than anything else, it was the expository work of the Bourbaki group that convinced them such a systematization could be achieved. Similarly, enormous progress has been made in recent years in the study of logical form or linguistic structure that should prove to be helpful to the linguist. Let us dub the study of these, and their role in the study of language generally, 'logico-linguistics' for want of a better term. Just as modern logic has developed with the help of both mathematicians and philosophers, so, presumably, logico-linguistics will develop fruitfully only with the cooperation of logicians and linguists. Such co-operations is an activity devoutly to be wished.

By the *structure* of a sentence let us mean simply its logical form, and by the structure of a text let us mean the sequence of logical forms of its constituent sentences taken in the same order. The sentences of a text "interanimate" each other in most intimate ways, as is well recognized, and this interanimation must be made explicit in the sequence of forms by suitable cross-referential devices. In Hiż's aletheism, for example, and in non-translational semantics generally, such devices are regarded as *sui generis* and are not made to rest upon some notion of direct reference independent of cross-reference (Hiż, 1969a). However, this may be, all manner of means for handling reference and cross-reference of whatever complexity must be at hand, if the study of logical form is really to get under way.

The question immediately arises as to just what a logical form is. Several criteria have been suggested elsewhere (see MARTIN, 1978a, Chapter XV). There are many variant ways of characterizing them, but all workers are agreed that they must be incorporated in a *system* very much in the manner in which some basic logic is always presupposed in the formalization of some area of mathematics or empirical science. It is very important to keep the system as simple as possible, both in linguistics and in the foundation of the sciences, lest complexities in the underlying logic be allowed to obfuscate rather than clarify. It is extraordinary methodological fact that the simpler the logic is required to be, the more is made to depend upon bringing out into the open *all* the non-logical predicates and terms determinative of the given subject-matter. There should be *no hidden predicates, no hidden variables or terms, no hidden quantifiers*, or the like. Complex logics—type-theory, set-theory, modal theory, model-theory, and the like—at best serve as heuristics and are not ultimately acceptable in their own terms. They posit too many hypothetical elements not required for the given subject-matter. As Einstein put it so beautifully: “If there are too many hypothetical elements one cannot believe one is on the right track. Thus [in the development of my work leading to the theory of general relativity] I came to *logical simplicity, a desperate man’s way to get on the right track.*”¹

What is probably the simplest known system of forms runs somewhat as follows. The usual *first-order logic* with identity is presupposed, with *virtual classes* and *relations*, extended to include a suitable version of Leśniewski’s *mereology* or theory of the part-whole relation between individuals. In addition, a theory of *events, states, acts, processes*, and the like, may be provided by the introduction of variables for them and of special *event-descriptive predicates*.

The explicit need for linguistic purposes of variables for events was perhaps first noted by C. S. Peirce, in his classic analysis of human acts of giving. “For instance”, he wrote (PEIRCE, 1931–58, 3.492), “A gives B to C may be represented by saying A is the first party [agent] in the [some?] transaction D, B is the subject [object] of D, C is the second party [patient] of D, [and] D is a giving by the first party of the subject to the second party.” The event-descriptive predicate here may be taken as

$$(1) \quad \langle A, G, B, C \rangle'$$

¹ In WHEELER (to appear). The italics are added.

in the context

$$(2) \quad \langle A, G, B, C \rangle D,$$

which expresses that D is an- A 's-giving- B -to- C transaction (or act or event). The expressions (1) and (2), however, contain hidden predicates that should be brought into the open, more particularly, predicates for the relations of being an *agent-of*, of being the *object-of*, and of being the *patient-of*.

Let ' e ', with or without primes or numerical subscripts, be variables hereafter for events, and let ' p Agent e ', ' p Patient e ', and ' x Object e ' express respectively that person p is the agent of e , p is the patient of e , and that the individual x is the object of e . And let ' $\langle G \rangle e$ ' express that e is an act of giving. (2) may then be given the more explicit form

$$(3) \quad (Ee)(A \text{ Agent } e \cdot \langle G \rangle e \cdot B \text{ Object } e \cdot C \text{ Patient } e).$$

Such a form contains no hidden material other than what is needed for the analysis of the constituent predicates.

In citing logical forms, it is customary to take the non-logical predicates and names at face value, and this practice will be followed here. Ultimately, of course, certain such predicates (and perhaps names) are to be taken as primitives, and thus characterized axiomatically, with the others then suitably defined. The whole array of predicates needed for a full natural language will of course turn out to be very extensive indeed, but not beyond human power to determine.

(3) is still far from adequate. An additional clause is needed to provide the appropriate *tense*. Also the handling of the preposition 'to' in 'A gives B to C' is left unaccounted for. Two obvious notions needed for tense are some deictic expressions '*now*' for the speaker's present moment and a temporal *before-than* relation, say B . Many (perhaps all) locutions involving tense, and also *aspect*, may be defined with these together with a part-whole relation between events. Mereology is thus needed in the theory of events as well as in that of individuals, and seems fundamental for the theory of aspect. An action is completed, for example—in many instances anyhow—just where all parts of it bear the relation B to *now*. Suppose ' e During *now*' expresses that e takes place during the *now* of the speaker. A conjunct in (3) may then be inserted to provide that the giving takes place *now*, thus providing for the present tense 'gives'. (3) is of course in the Fregean tense of timelessness; in fact, all of its conjuncts are.

The preposition 'to' in 'A gives B to C' seems best handled by the to relation of patency. Let ' $To_{Patient}$ ' symbolize this where ' $e To_{Patient} p$ ' expresses that e is an action taking place with p as patient. The full logical form for 'A gives B to C' then becomes

- (4) '(Ee)(A Agent e . $\langle G \rangle e.e$ During now.B Object $e \cdot e$ To_{Patient} C)'.

Here of course the English 'give' is "represented" by 'G', the final 's' in 'gives' by the deictic 'now', and the English 'to' by 'To_{Patient}'. The whole then represents—in Peirce's meaning of the word, notice—the tensed English sentence 'A gives B to C', where 'A', 'B', and 'C' are taken as proper names [say, 'Adam', 'Belchamber' (by Howard Sturgis), and 'Cathy']. Note the harmless ambiguity here of 'B'.

Because of the presence of 'to' in the English 'A gives B to C' let us no longer write just ' $\langle A, G, B, C \rangle e$ ' as above, but instead

- (5) ‘⟨A, G, B, To, C⟩e’, regarded now as short for
 ‘(A Agent e · ⟨G⟩e · B Object e · e To_{Patient} C)’.

Note that the definiens here, and hence of course the definiendum, is tenseless, tense and aspect being provided by additional clauses. This definition is already in effect a rule of transformation, of which we shall speak in a moment.

Provision should also be made, in the system of logical forms, for *semiotical* items in the form of a systematic, inscriptive *syntax*, *semantics* (both extensional and intensional), and *pragmatics*. The presence of these is needed, not only to handle sentences that are at first blush metalinguistic, but also to bring to light hidden referential and other features of sentences that at first blush are not. Thus there is no clear line of demarcation between object-language and metalinguistic sentences, and hence it is of little interest to separate off two parts of the system, one dealing with one kind, the other with the other. More important is the *unification* of the two. Every natural language contains its own hierarchy of metalanguages, each metalanguage in turn being a sublanguage of its metalanguage. No interest attaches to a language at any one level more than at any other, and a logical form may end up being located at any level depending upon its internal features. Some level higher than the lowest is usually needed for the handling of a text, with cross-references between items in its various sentences. Especially simple examples are of pronouns used referentially or cross-referentially (either anaphorically or epiphorically).

Consider next the text

'Adam gives *Belchamber* to Cathy. He loves her.'

The 'he' and 'her' here are of course used cross-referentially by the speaker to 'Adam' and 'Cathy', respectively. The form for 'He loves her' will thus not only contain deictic expressions '*he*' and '*her*' but also a clause to the effect that the speaker takes these words in the context to have the desired cross-referentiality to the respective words as occurring in the preceding sentence. The full spelling out of the logic of cross-reference is a much more complicated matter than at first appears. (Cf. HİZ, 1969b; also MARTIN, 1978b, Chapters VII and VIII, and SMABY, 1971.) It is the sort of matter we know all about until asked for an explicit formulation. In particular, it must presuppose a full theory of *inscriptions* or sign-events, in terms of which deictic expressions in general may be handled. Logicians tend not to like the complexities of inscriptions and of the semantics and pragmatics based on them. Their role in language seems so very basic, however, as to be unavoidable. They constitute the very bedrock of language, all of our linguistic activity consisting of writing them, uttering them, questioning them, doubting them, and so on and on. There would thus seem to be no adequate way at all of avoiding the explicit introduction of inscriptions as values for the expressional variables. On the contrary, they should be welcomed with open arms.

Note that the logical form (4) is a conjunction, each conjunct of which brings out some basic structural feature of the English original. One is the feature of agency, as noted, one concerns the nature of the act or event, one the tense, one the object, and one the patency. More complex sentences will of course need more conjuncts, and further kinds of sentences will need further kinds of conjuncts. The first task of logico-linguistics, as conceived here, is to provide the full array of structural relations needed for logical forms of sentences and of full texts of any complexity. Of course "representations" of the non-logical (non-semiotical, non-grammatical) words occurring are assumed available. The logical form then consists of these representations occurring in the very order in which their originals occur in the original sentence, interspersed, prefixed, or suffixed with suitable logical and structural material. The interspersed material spells out in full detail the structural roles, so to speak, of the representatives in the given context.

The logical form is thus the "meaning" of the original, its semantic or pragmatic structure, if you will. (Cf. MARTIN, b, to appear.) All structural ambiguities are assumed to be disambiguated. If an English word is itself

ambiguous, it will have as many distinct representatives as are required. Ultimately each and every unambiguous word of the natural language must be provided its appropriate "logic". Perhaps some can be listed as primitives, the others then being definable, with suitable meaning-postulates given for the primitives, as already suggested.

The total vocabulary required for the theory of logical form may be summarized then roughly as follows: representatives of non-logical predicates and names; some specific logico-grammatical predicates for agency, patency, objectuality, and so on; suitable primitives for syntax, semantics, and pragmatics; 'P' for the part-whole relation between individuals and between events; 'B' for the temporal before-than relation; deictic expressions such as 'now', 'here', the demonstrative 'that', 'he', 'she', and so on; a suitable notation for handling events, states, acts, and so on; variables for individuals, events, persons, and inscriptions with quantifiers upon them; and a notation for handling virtual-classes and relations of or between or among all these various kinds of entities. Concerning all the primitives, of course, suitable axioms or meaning-postulates are to be laid down.

A good deal is made in this present treatment of the theory of *gerundives*, which seems not to have been given an exact logical characterization. The half-diamond braces in ' $\langle G \rangle$ ' have the effect if an operator on G or ' G ' yielding the gerundive ' G -ing'. Where ' G ' represents 'give', ' $\langle G \rangle$ ' represents 'giving', so that

'Giving e ' may be defined as ' $\langle G \rangle e$ '.

All manner of complex gerundive constructions may be handled by this notation. Thus,

- (6) 'John's phoning leading to his going led to Adam's
giving *Belchamber* to Cathy'

may be given the approximative form

- (6') '(Ee_1)(Ee_2)(Ee_3)(Ee_4)(Ee_5)($\langle J, Ph \rangle e_1 \cdot \langle e_1, L, e_2 \rangle e_3 \cdot \langle J, Go \rangle e_2 \cdot \langle e_3, L, e_4 \rangle e_5 \cdot e_5 B_{now} \cdot \langle A, G, B, To, C \rangle e_4$)',

with the obvious symbolization. And this is readily distinguished from

- (7) 'John's phoning led to his going's leading to Adam's
giving *Belchamber* to Cathy',

with the approximative form

$$(7') \quad (\text{E}e_1) \dots (\text{E}e_5)(\langle J, \text{Ph} \rangle e_1 \cdot \langle e_1, L, e_5 \rangle e_3 \cdot e_3 \text{Bnow} \cdot \langle J, \text{Go} \rangle e_2 \cdot \langle e_2, L, e_4 \rangle e_5 \cdot \langle A, G, B, \text{To}, C \rangle e_4)^2$$

In both of these forms the presence of 'his' as cross-referential to 'John' in the original is left unaccounted for. Also complex event-descriptive predicates are used that are further reducible. Thus

' $\langle J, \text{Ph} \rangle e_1$ ' may be regarded as short for ' $(\langle \text{Ph} \rangle e_1 \cdot J \text{ Agent } e_1)$ '.

Thus to say that e_1 is a John's-phoning act or event is to say that e_1 is a phoning of which John is the agent or doer. Note that the apostrophe and final 's' in 'John's' in this context is a suffix for agency—and thus definable in terms of the by-relation of agency—and not for possession or ownership 'John's' in this sense contrasts of course with the use of 'John's' in 'John's necktie', which utilizes the apostrophe-'s' suffix, definable in terms of the of-relation possession. The full spelling out of the cross-referentiality of 'his' to 'John' is of course somewhat complicated. (Cf., however, MARTIN, 1978c.)

Note that the forms (6') and (7') enable us to see in a very simple way the *logical consequences* of the original sentences. Its logical consequences are of course of primary importance to the structure or meaning of a sentence or text. The conjunctive forms, covered by existential quantifiers, lead in a very direct way to many of the logical consequences that must be provided for, in view of the general quantificational principle that

$$\vdash \Gamma (\text{E}e_1) \dots (\text{E}e_n)(A_1 \cdot \dots \cdot A_k) \supset (\text{E}e_{n_1}) \dots (\text{E}e_{n_j})(A_{k_1} \cdot \dots \cdot A_{k_m}) \square,$$

where each of A_{k_1}, \dots, A_{k_m} is one of A_1, \dots, A_k and each of e_{n_1}, \dots, e_{n_j} is one of e_1, \dots, e_n .

Thus it is a consequence of (6') that John phoned, that John went (somewhere), that Adam gave *Belchamber* to Cathy, and so on. Still further consequences are provided for, of course, by taking into account the meaning-postulates governing the non-logical predicates. It then follows also from (6') that John did something, for example, in view of the meaning-postulate that

$$(e)(\langle \text{Ph} \rangle e \supset \langle \text{Do} \rangle e),$$

* These examples are adapted from similar ones in HARRIS (1976).

that every phoning is a doing. The importance of the notion of logical consequence in the study of logical form and meaning cannot be over-emphasized, as has frequently pointed out.

As further examples of meaning-postulates concerning 'G'—assuming it to be a primitive—we would have a limitation law, that

$$(e)(\langle G \rangle e \supset (Ep)(Ex)(Eq)(p \text{ Agent } e.x \text{ Object } e.q \text{ Patient } e)),$$

as well as that

$$(e)(\langle G \rangle e \supset \langle Do \rangle e).$$

The converse of this does not obtain. Not all doings are givings. Likewise the converse of the limitation law does not obtain. For example, acts or states of *owing* are such as to have an agent, an object, and a patient but are not therewith acts of giving. Many meaning-postulates will be such as to incorporate a doctrine of *semantical categories* or of intersignificance. Only certain kinds of individuals are suitable to be objects of acts of giving, for example. One cannot give the moon to anyone, except perhaps metaphorically.

Nothing has been said thus far about intentional—or intensional—contexts. The method used here for handling these is an adaptation of Frege's *Art des Gegebenseins*, or mode of linguistic description (see FREGE, 1879, § 8, and FREGE, 1967, 2nd paragraph). We could let '*e* Under *a*' express primitively that *e* is taken under the predicate-description *a*, i.e., that *e* is regarded as having the predicate *a* apply to it. A better form, however, would be the pragmatized one,

$$'p \text{ Under } e, a',$$

to the effect that person (speaker or hearer) *p* takes *e* under *a*. The form

$$\langle p, \text{Under}, e, a \rangle e'$$

then enables us to express that *e* is an *act* of *p*'s taking *e* under *a*. A still more useful intentional relation, for some purposes, is the relation *That* in contexts of indirect discourse. Suppose *e* is an act of believing and that *a* is the sentence giving the *content* of what is said to be believed. The relation here between *e* and *a* may be symbolized by '*That*_{Content}', standing for the that-relation of content. Consider, by way of an example,

'John believes that Adam gave *Belchamber* to *Cathy*'.

A form for this is forthcoming as follows.

- '(Ee)(Ea)(J Agent e.⟨Blv⟩e · e During now · e That_{Content} a
- '·(Ee')(A Agent e' · ⟨G⟩e' · e' B now · B Object e' · e' To_{Patient} C)'a'.

Note that, 'a' here being a variable for inscriptions, the clause "—'a'" states merely that a is an inscription of the appropriate shape '—'.

Given a natural-language sentence or text, how do we arrive at a logical form for it? Any attempt to answer this question leads at once to the topic of rules of transformation. Given the parent English sentence

- (8) 'Adam gives Belchamber to Cathy',

how do we arrive at (4), other than by what has been called "mysterious translation"? And conversely, given (4), how can we make good English sense of it by arriving at (8)? Let us consider these questions one by one.

Let us consider first (4) and the steps required to enable us to gain (8), its English original. From (4), by means of the definition (5) we gain

- (4.1) '⟨Ee⟩(⟨A, G, B, To, C⟩e · e During now)'.

The definition (5) may be referred to as the *Rule of Compound-Predicate Introduction* and *Elimination*. Replacement of definiens by definiendum is clearly the introduction of a compound predicate, and replacement of definiendum by definiens, its elimination.

Next we replace the predicate 'G' in (4.1) by the tenseless English 'give' (taken in the appropriate meaning) of which it is representative. This is justified by the *Rule of Representation for Predicates*. The results of this transformation, going either way, are regarded as logically equivalent. Thus we gain

- (4.2) '⟨Ee⟩(⟨A, give, B, To, Cathy⟩e · e During now)'

from (4.1).

Strictly we should write 'To_{Patient}' throughout for 'To'. This is legitimized, however, by a *Rule of Subscript Introduction (Elimination)* for such subscripts, such subscripts normally being dropped in ordinary language.

We need now a logical *Abstraction Principle* as follows. Let { $p'x'Qq'\exists-p'-x'-Q-q'-$ } be a pentadic virtual relation among persons p' , individuals x' , dyadic virtual relations Q , and persons q such that $-p'-x'-Q-q'-$,

where ' $\neg p'x'Qq'$ ' is a suitable sentential form. The principle now needed is that

$$\vdash p\{p'x'Qq' \exists (\text{Ee})(\langle p', G, x', Q, q' \rangle e \cdot e \text{ During } now)\}x, \text{To}, q \equiv \\ (\text{Ee})(\langle p, G, x, \text{To}, q \rangle e \cdot e \text{ During } now).$$

This is a familiar enough kind of logical principle in the theory of virtual relations.

A definition of 'gives' in the present tense is needed in terms of the timeless 'give'. But clearly

'gives' may now abbreviate

$$\{p'x'Qq' \exists (\text{Ee})(\langle p', \text{give}, x', Q, q' \rangle e \cdot e \text{ During } now)\}.$$

This definition is *Present-Tense Introduction* (and *Elimination*). By means of this definition and the Abstraction Principle just given, we gain

$$(4.3) \quad 'A \text{ gives } B \text{ To } C'$$

from (4.2). But clearly the relational predicate 'To' is the representative of the English 'to'. Hence by the Rule of Representation for Predicates, we gain

$$(4.4) \quad 'A \text{ gives } B \text{ to } C'.$$

And finally, by a *Rule of Representation for Names*, we gain (8) itself. Thus (8) is gained by derivation from (4).

Note that the Rules are such as to allow also the converse derivation, of (4) from (8). Note also that the Rules are suitably restricted so as not to allow deviant derivation of either an incorrect logical form, on the one hand, or English gibberish, on the other. In particular the Rule of Compound-Predicate Introduction and Elimination, definition (5), is highly restricted. The Abstraction Principle used is merely an instance of a much more general principle, and incorporates some features of what is essentially Reichenbach's "event-splitting". (See REICHENBACH, 1947, p. 271.) The definition (9), Present-Tense Introduction and Elimination, is restricted to just the predicate 'give' and to "subjects" in the third-person singular, so to speak. Note also the "naturalness" of these rules; they are either more or less standard rules in the kind of logical framework presupposed, or are natural extensions of it to provide for the inclusion of words of natural language in the normal ways in which they are allowed to occur.³

³ Cf. also MARTIN (a), in which, however, triadic relations were not considered. The transformations provided above may thus be the first yet given involving triadic relations

There are no doubt many predicates in English other than 'give' for which essentially the same derivations can be given *mutatis mutandis*.

'G' for 'give' is a triadic relation. Sentences related to (4.3) are of course

- 'A gives (to) C B'
- 'B is (being) given by A to C'
- 'B is (being) given to C by A'
- 'C is (being) given, by A, B'

and

- 'C is (being) given B by A'.

There are just these five, triadic relations having just five converses. Logical forms for all of these sentences may readily be supplied, and derivations for them given. The problem then remains of showing precisely how logical forms for these sentences are interrelated, all conceivable differences of nuance being taken into account.

Let us now have a closer look at (6), and perhaps at (7) also. (6') was said to give an approximative form for (6). In (6') there are existential quantifiers for the separate acts of John's phoning, of his going, and of Adam's giving *Belchamber* to Cathy. Actually, however, these acts are clearly intended to be unique. It is some one and only one act of John's phoning, of John's going, and of Adam's giving *Belchamber* to Cathy that are under discussion. Hence they should be handled by definite Russellian descriptions, some additional information being presupposed sufficient to render them unique.

Note that (6') and (7') contrast sharply in this regard with (4). The existential quantifier in (4) provides for at least one act of Adam's giving *Belchamber* to Cathy, leaving it open that he might give it to her several times, or even give her several copies on as many occasions. '*Belchamber*', or 'B' as occurring in (4), might be construed as the name of some one copy of that novel. In some other occurrence, it could be construed as a name for some other copy. No matter, the existential quantifier involved covers at least one act of the appropriate kind, perhaps more than one.

For a closer handling of (6) we need then a description

$$'(i e_1 (\langle J, Ph \rangle e_1 \cdot F_0 e_1))'$$

in which the presence of prepositions in the English readings is really taken seriously
The extension to quadratic relations is of course immediate.

where ' $F_0 e$ ' specifies the conditions presupposed assuring that one and only one act of John's phoning is under discussion. And similarly for

$$'(i e_2(\langle J, Go \rangle e_2 \cdot G_0 e_2))'$$

and

$$'(i e_4(\langle A, G, B, To, C \rangle e_4 \cdot H_0 e_4))'$$

for John's going and for Adam's giving *Belchamber* to Cathy. And similarly also for the leading. It is one event of John's phoning leading to his going that led to Adam's giving *Belchamber* to Cathy. Hence we need a fourth description

$$'(i e_3 \langle (i e_1(\langle J, Ph \rangle e_1 \cdot F_0 e_1)), L, (i e_2(\langle J, G_0 \rangle e_2 \cdot G_0 e_2)) \rangle e_3)'$$

for the one's leading to the other. No additional clause for uniqueness seems needed here; presumably it is provable from the uniqueness of the two acts. (Similarly the second leading is unique also, but nothing in (6) demands this, so that no description need be introduced for it.) To save writing, let the four foregoing descriptions be abbreviated by ' E_1 ', ' E_2 ', ' E_4 ', and ' E_3 ', respectively. A better form for (6) is then

$$(Ee_5)(\langle E_3, L, E_4 \rangle e_5 \cdot e_5 \cdot B \text{ now}). \quad (6'')$$

Let us consider now the transformations needed to derive (6) from (6'') and conversely. Recall that

$$\langle J, Ph \rangle e_1 \text{ is short for } (\langle Ph \rangle e_1 \cdot J \text{ Agent } e_1).$$

Also

$$'e_1 \text{ By}_{\text{Agent}} J' \text{ may be taken as short for } 'J \text{ Agent } e_1',$$

and

$$'Ph-ing e_1' \text{ as short for } '\langle Ph \rangle e_1'.$$

This last definition is the transformation of *Gerundive Introduction* (and *Elimination*). The definition of

$$'J's Ph-ing e_1' \text{ as } '(Ph-ing e_1 \cdot e_1 \text{ By}_{\text{Agent}} J)'$$

is the transformation of *Apostrophe and Final 's' Introduction* (and *Elimination*) for *Agency*. Now, where uniqueness is assured, we may define

$$'J's Ph-ing' \text{ as } '(i e_1(J's Ph-ing e_1 \cdot F_0 e_1))',$$

and similarly

$$\text{'his Go-ing' as } (\iota e_2(\text{his Go-ing } G_0 e_2)).$$

These transformations are *Agentive Gerundive Introduction* (and *Elimination*). Here of course we are considering these transformations only in the special cases needed. In view of these definitions it is provable that

$$J's Ph-ing = E_1$$

and

$$\text{his Go-ing} = E_2,$$

subject to appropriate scope-indicators for the descriptions (as in *14, *Principia Mathematica*). In a somewhat similar way

$$E_1's L-ing E_2'$$

may be defined, it then being provable that

$$E_1's L-ing E_2 = E_3.$$

And similarly

$$A's G-ing B To C = E_4.$$

Now from (6'') we gain

$$(Ee_5)(\langle E_1's L-ing E_2, \text{lead}, E_4 \rangle e_5 \cdot e_5 B \text{ now}),$$

by replacement of identities (subject to proper scopes for the descriptions) and the Rule of Representation for Predicates. By an appropriate Abstraction Principle and *Past-Tense Introduction* transformation, we gain then

$$E_1's L-ing (to) E_2 \text{ led (to) } E_4,$$

and from this to

$$J's Ph-ing L-ing (to) his Go-ing led (to) A's G-ing B To C.$$

By the Rules for Representation for Predicates and Names, we then gain (6) itself. And conversely, from (6) back again to (6'').

There is still more to do, but this much must suffice for the present. The presence of 'to' in 'leading to' and 'led to' is unaccounted for. Also the cross-referentiality of 'his' to 'John's' is left unanalyzed.

Harris considers the example

'John's phoning leading to Frank's arrival prevented our escape'.

This sentence cannot be handled like the foregoing, lest the existence of the very act prevented follow as a logical consequence. Prevention in fact is to be handled intensionally, where

' $e_1 \text{ Prvt } e_2, a$ '

expresses that e_1 prevents (timelessly) e_2 as taken under the predicate-description a . Where N is the null event, a meaning-postulate for 'Prvt' is then that

$(a)(e_1)(e_2)(e_1 \text{ Prvt } e_2, a \supset e_2 = N)$.

But of course the null event may be taken under all manner of *Arten des Gegebenseins*.

If any of the sentences considered are embedded in a context in which their "meaning" deviates from the results of the kind of analysis given, the sentence then is to be reconstructed or "reread" accordingly. Strictly no sentence should be considered in isolation, as already remarked. However, no harm can arise from this provided a suitable *ceteris paribus* clause obtains. And, of course, analysis must get started at some point with disambiguated sentences.

These examples of derivations are, to be sure, merely illustrative and not intended as an outline of a general method. The question naturally arises as to how far this method can be developed. Of course, one cannot say in advance, but the horizon seems unlimited. Logico-linguistics is still in its infancy and many approaches will have to be made.

At some point in the study of logical form a relation of reference will be needed. For various reasons it seems best to provide for this pragmatically, i.e., by means of a form in which the user of language, speaker or hearer, is explicitly brought in. Thus we may let

' $p \text{ Ref } a, x, b$ '

express that p uses the inscription a as occurring in b to refer to individual or event x . The use of this form of locution is thought sufficient for saying whatever is needed about reference, and indeed cross-reference as well.

Another fundamental logico-linguistic notion is that of *paraphrase*. One way of construing this is semantically in terms of logical consequence, as with Hiż. (Cf. especially Hiż, 1964, and 1968.) A sentence a is a para-

phrase of b then just where a and b are logical consequences of each other. Alternatively, one might require some strong logical equivalence or synonymy rather than mutual consequence. Both notions being semantical, it would seem that the theory of paraphrase, as construed in either way, would already be contained in the theory of transformations. Not, however, if paraphrase were to be construed pragmatically. Let

(10) ' p Prphrs a, b '

express that speaker or user p paraphrases a as b , where a and b are inscriptions for sentences. Use of this form, together with further notions definable in terms of it, would give us great pliability in dealing with particular users' linguistic behavior, with idiolects and dialects, and in fact would help to pave the way for socio-linguistic study in general. And of course the linguistic behavior of some users might be such as to accord with either of the semantical uses of 'paraphrase'. Some users might in fact paraphrase a as b just where a and b are synonymous in view of the theory of transformations. Or paraphrase might be allowed to hover somewhere between synonymy and mutual consequence. The use of the pragmatic form thus in no way excludes such considerations concerning paraphrase as might rest on purely semantical features.

It has already been remarked that a natural language is in effect the underlying logic in a kind of notational disguise, with natural words in place of the predicates and names of logic and with deletion of the logico-semiotical material needed in the logical forms. By 'logic' here of course one means logic-cum-semiotics-cum-event-theory as sketched above. The relation of logic to language, in fact, is rather similar to that of modern set-theory to numerical mathematics proper. Just as the resources of set-theory vastly exceed what is needed for, say, the theory of functions of a complex variable, so the resources of logic exceed what is needed for natural language. Language and mathematics are respectively the tips of the icebergs, the greater portions of which float hidden beneath the surface.

Transformation rules enable us to "derive" natural-language sentences from corresponding logical forms and conversely. We have seen from the examples above that these are either mere definitions or "natural" rules of a logical kind. Still further kinds of transformation may or may not turn out to be needed. The problem, however, of transforming formulae of set-theory into straight numerical formulae, and conversely, scarcely arises for anyone familiar with set-theory. The very meanings of the

numerical formulae are given by their set-theoretic representations, all proofs of them rest ultimately upon the set-theoretic axioms, and all definitions of numerical and functional notions go back ultimately to the set-theoretic primitive or primitives. In view of these considerations, we may envisage that the day will come when the logico-linguist will pass as easily from a natural sentence to its logical form and back again as a set-theorist does now between numerical and corresponding set-theoretic formulae.

Set-theory, in its modern rigorous form, is unthinkable without the underlying truth-functions and quantifiers. It is thus the result of a joint venture between mathematicians and logicians. In a similar vein logico-linguistics, as envisaged here, can best be furthered cooperatively, as already suggested. The task of the logician is primarily to supply the system of logical forms with such rules of transformation as are logical in character. The linguist then becomes the guardian of the definitions admitted and of the restrictions required in the other rules. These all run a hazardous course between being too narrow or too general, too narrow if they cover merely a single case (which can usually be covered by an *ad hoc* restriction), and too general if they fail to lead always to (or from) natural sentences. But the logician can help also in formulating definitions just as the aid of the linguist is indispensable in characterizing such notions as agency, patency, and so on. What is now needed in logico-linguistics is a vast, co-authored, three-volume work—a veritable *Principia Linguistica*.

The importance of the definitional transformations needs emphasis. The logical ones for the most part are already provided in the underlying logic. The very heart of the material intrinsic to linguistics is thus couched in the definitions. Whitehead was perhaps the first to have noted this important methodological point when he commented that “the act... [of giving a definition within a formal system] is in fact the act of choosing the various complex ideas which are to be the special object of study. The whole subject depends upon such a choice.” Thus they “are at once seen to be the most important part of the study.” (WHITEHEAD, 1906, p. 2.) Although definitions are often regarded as mere conventions of notational abbreviation, their very fundamental role belies this fact and gives some reason for preferring the Polish view in which definitions are regarded, not as abbreviations, but as axiomatic equivalences. The difference is akin to the ancient one between regarding definitions as real or as nominal.

In addition to gaining a proper array of definitions, either real or nominal, much effort must be expended in developing the theory of agency, patency, and so on, including the full list of relations needed for expressing what is needed concerning "who does something, who experiences something, who benefits from something, where [and when] something happens, what it is that changes, what it is that moves, where it starts out, and where it ends up..." (FILLMORE, 1968.) The problems here are not overwhelming if only a real effort were made to deal with them systematically.⁴ The theory of these is to structural and transformational linguistics what the theory of the membership-relation or -relations is to set-theory. But, as has often been observed, set-theory is easy as compared with the complexities to be faced in the study of the structure of language. Also there is no hope for a radical reducibility in this kind of study to that of only a few primitives, as in set-theory in which all notions are ultimately definable in terms of the primitive membership-relation(s). *Mathematica brevis, linguistica longa*, so to say.

It is interesting to ask what, in a logico-linguistic system, corresponds with theorems, in a set-theoretic one. Of course, the axioms characterizing agency, and so on, together with meaning-postulates governing the non-logical predicates and terms correspond more or less roughly with the set-theoretic axioms. And of course there are interesting theorems forthcoming from such axioms, and some of these theorems will make use of definitions. Among these latter are theorems stating the equivalence of parent natural-language sentences and texts with their logical forms and sequences of such, such theorems providing a full meaning-analysis of the sentences and texts involved. If the main task of the study of language is thought to be a full meaning-analysis, then such equivalence-theorems are to be regarded as providing the main content of the subject.

Another point. All working mathematics may be incorporated in the set-theoretic framework, with of course suitably powerful axioms such as the *Ersetzungssaxiom*. Likewise it is to be presumed that whatever is of interest in "working" grammar, or perhaps even traditional grammar, will find its proper place somewhere along the way in the maelstrom of transformations needed here. If this should turn out not to be the case, and that some items of independent grammatical interest and worth preserving cannot be accommodated, suitable extensions will of course be required.

⁴ Essentially this point was made by Kurt Gödel to the author in a conversation in March, 1976.

It might have been advisable throughout not to have used special letters as "representatives" of English words, but to have used instead those very words themselves as logical symbols. Such usage would have been of help in visualizing how closely logic and natural language are in fact interrelated. The argument against such usage, however, is that some one natural language would have then been picked out as in some way fundamental. Of course, English has been the favored language here, but this is a mere historic accident. The system of English requires one set of transformation rules to "generate" its permitted forms, other languages, other sets, and perhaps even other kinds. The underlying logical substructure, however, is thought to be the same for all languages. The underlying system provides not a "universal grammar" but a universal logic, much of the grammar being couched in the peculiarities of the transformations required for a particular natural language. The considerations here thus lend no support to the thesis of a universal grammar, but do rest fundamentally on the thesis of a universal logic. In similar vein a set-theory may be regarded as a kind of universal mathematics. All areas of working mathematics, all sublanguages of mathematics, are contained in the overall system—all kinds of algebra, of topology, of geometry, all kinds of numbers, of variant structures of mathematical interest, and so on. The thesis of a universal logic for all languages seems thus on a par with the contention that set-theory provides a kind of universal mathematics.

The problem of conceptual change has often been raised against the synchronic kinds of considerations here. Linguistic change might be thought to be beyond the purview of these methods. A diachronic characterization of language is needed, we are told. Very well, let us provide it; the pragmatic resources are at hand. Diachronic considerations concerning language are not so very different from those needed in the sophisticated discussion of method in the sciences generally. Tisza's "dynamics of logical systems" is an item in point, of especial interest to the methodology of contemporary physics. (TISZA, 1963.) Each system may be thought to characterize basic features of a language throughout a given time-span. When the language changes radically enough, the one system will give way to another better characterizing the changed features. The study of language-change, and even of the history of languages, poses no objection to the feasibility of the methods here.

Condillac's famous observation that "une science n'est qu'une langue bien faite" is thus as applicable to linguistics as it is to mathematics or

physics. As structural linguistics grows to maturity, it too will no doubt wish to aim more and more towards the kind of logical perfection first envisaged by Frege a century ago.

References

- FILLMORE, Charles F., 1968, *Lexical entries for verbs*, Foundations of Language, vol. 4, pp. 333–393
- FREGE, G., 1879, *Begriffsschrift* (Halle)
- FREGE, G., 1967, *Über Sinn und Bedeutung*, in: Kleine Schriften (Darmstadt)
- HARRIS, Zellig, 1976, *On a theory of language*, The Journal of Philosophy, vol. LXXIII, pp. 253–276
- Hiż, H., 1964, *The role of paraphrase in grammar*, Monograph Series on Languages, and Linguistics, vol. 17, pp. 97 ff.
- Hiż, H., 1968, *Computable and uncomputable elements of syntax*, in: Logic, Methodology, and Philosophy of Science, vol. III (North-Holland Publishing Co., Amsterdam)
- Hiż, H., 1969a, *Alethic semantic theory*, The Philosophical Forum, vol. 1, pp. 438–451
- Hiż, H., 1969b, *Referentials*, Semiotica, vol. II
- MARTIN, R. M., 1978a, *Semiotics and linguistic structure* (The State University of New York Press, Albany)
- MARTIN, R. M., 1978b, *Events, reference, and logical form* (The Catholic University of America Press, Washington)
- MARTIN, R. M., 1978c, *On Carnap's semantics, Hiż's notion of consequence, and deep structure*, in: MARTIN (1978b)
- MARTIN, R. M., a, *Some protolinguistic transformations*, in: Pragmatics, truth, and language, Boston Studies in the Philosophy of Science (D. Reidel Publishing Company, Dordrecht)
- MARTIN, R. M., b, *On meaning, protomathematics, and the philosophy of nature*, in a volume devoted to Moritz Schlick, ed. E. Gadol (to appear)
- PEIRCE, C. S., 1931–58, *Collected papers*
- REICHENBACH, H., 1947, *Elements of symbolic logic* (The Macmillan Co., New York)
- SMABY, R. M., 1971, *Paraphrase grammars* (D. Reidel Publishing Company, Dordrecht)
- TISZA, Laszlo, 1963, *The logical structure of physics*, in: Proceedings of the Boston Colloquium for the Philosophy of Science, 1961–62, Boston Studies in the Philosophy of Science (D. Reidel Publishing Company, Dordrecht), pp. 55–71
- WHEELER, John Archibald, collection: *Mercer Street and other memories*, in: Albert Einstein 1879–1979, eds. P. C. Aichelburg and R. U. Sexl (Vieweg-Verlag, to appear)
- WHITEHEAD, A. N., 1906, *The axioms of projective geometry* (Cambridge University Press, Cambridge)

SOCIOLINGUISTIC METHOD AND LINGUISTIC THEORY

DAVID SANKOFF

Université de Montréal, Montréal, Québec, Canada

1. Introduction

What we may call ‘the sociolinguistic method’ is neither new to socio-linguistics, nor universally adhered to by sociolinguists, nor—strictly speaking—a method. It is basically a working hypothesis with a distinctive (within linguistics) methodological and conceptual apparatus, built up over the last ten to fifteen years in response to the particular needs of research guided by this hypothesis. The essential point is that the primary data for the study of linguistic structure, function, and change is the spoken language, more specifically spontaneous unreflecting speech in its natural context. This principle is best exemplified by the research of William Labov and associates (LABOV, 1963, 1966, 1972a, 1972b, LABOV *et al.*, 1968, 1972, 1977, 1978), though the present paper is based on the work of the Montreal French Project (e.g. SANKOFF, 1978; THIBAULT, 1979; G. SANKOFF, 1980). We shall trace some of the consequences of the natural speech hypothesis, not only on the methodological level, but also extending into linguistic theory. On the theoretical level we focus on two debates: one about the existence of the ‘syntactic variable’ and the second on the ‘wave’ theory.

2. The sociolinguistic paradigm

First, we stress how radically different for linguistics is this preoccupation with natural speech. The primary source of data for much modern linguistics, especially syntax, is the introspection of linguists themselves about grammaticality, meaningfulness and paraphrase. The other major source, especially in phonology, is work with informants using classical

elicitation procedures. The closest approach to the study of natural speech might be the transcription and editing, with the help of informants, of narrative elicited in a rather formal context. All of these types of data bear not so much on the language but rather on what someone thinks about the language.

One might take the point of view that any distortions this introduces are second-order effects—that the generalization to be drawn from one type of data would be much the same as from the other, and this is no doubt true in some situations for some purposes. It does not seem to be true, however, for the study of urban speech communities favoured by sociolinguists, loci of the confrontation of standard and non-standard speech varieties, and dominant versus minority languages, where ideas about who says what and how, in which contexts and for what purposes, are greatly affected by a linguistic ideology propagated by the schools, the media and other institutions (BOURDIEU and BOLTANSKI, 1975), and generated in the social praxis of a society characterized by class inequality and ethnic division. An informant's opinions and judgments reflect, in unknown proportions, 'true' linguistic structure, unrealistic norms, and false stereotypes. When the linguist doubles as informant, we can add theoretical prejudice, conscious or unconscious, to the list of factors intervening between linguistic structure and speaker's intuition.

In any case, one of the most important consequences of the natural speech approach is not that it gets the 'right' answer where another method gets the wrong one, but that it provokes different types of questions to be answered, and hence prompts theorizing about language with focus and scope distinct from linguistics based on introspection or elicitation.

On the methodological side, the most obvious characteristic which distinguishes this kind of research is the importance it places on counting, quantification and statistics, none of which play a role in more traditional approaches to phonology, morphology, syntax and semantics. When a corpus of natural speech is studied, whether produced by one or several speakers in one or different contexts, it is not only the contrast between what is said versus what does not seem to be said which must be accounted for, but also that certain elements occur more frequently than others and that these frequencies, not just presence and absence, are systematically conditioned by co-occurring quantities in the phonological or syntactic environment. These are striking facts of a linguistic nature, systematic and suggestive of a range of generalizations. The methodology for accurately describing these facts and the theoretical apparatus for explaining

them does not exist within paradigms which depend on introspection or elicited data.

Note that despite its assumptions about recurring events and patterns, and its reliance on counting and quantitative analysis, our approach cannot be characterized as any sort of strict positivism. Indeed, there is a debate within the field as to the appropriate domains of interpretation versus 'distributionalism', which is essentially a type of positivism. Further, work in this field inherently constitutes a critique of existing social institutions connected with language since it inevitably unmasks and demystifies oppressive ideology about popular or minority speech varieties. (I would not claim the same for more psycholinguistic approaches to the study of the spoken language.)

3. The demise of free variation

The developments we shall discuss form a logical progression, not always corresponding strictly to their chronological emergence but permitting a more coherent exposition.

We may start with the notion of free variation in traditional phonology. Given a number of allophones, once all contextual factors, such as the surrounding phonological environment of a segment, stress, relation to word and syllable boundary, and morphological status, have been examined for whether or not they determine, singly or in combination, the choice of an allophone, if some contexts remain in which two different allophones may both be realized, they were said to vary freely. Free variation was considered relatively rare, transitory, linguistically unimportant, and to reflect a certain pathology, if not in the language itself, at least in the linguistic analysis.

This concept was one of the first casualties of the sociolinguistic study of natural speech. First, rather than a pathological rarity, the alternation of variants even in the most carefully defined contexts, is widespread, both in terms of the number of phonemes subject to variation and the range of contexts in which it occurs. Second, the variation is not 'free', but highly structured, even if this structure is quantitative rather than the qualitative or categorical conditioning familiar in linguistics. For example, instead of 'allophone *A* in context *X* and allophone *B* in context *Y*', we have 'variant *A* occurs 80% of the time in context *X* but only 30% of the time in context *Y*'. Instead of 'feature *R* determines that allophone *A* be realized whereas feature *Q* excludes it', we have 'feature *R*

favors variant *A* and *Q* disfavors it'. Indeed, many of the quantitative results of this type are more subtle and accurate versions of qualitative claims of traditional phonological treatments, claims which suffer from 'categorical perception', a type of bias whereby certain variable tendencies are interpreted as categorical when intuition or unsystematic impressions are relied upon.

A third aspect of the notion of free variation has to do with its alleged unimportance for linguistic theory. On the contrary, in any theory accounting for natural speech use, the study of variation becomes the key to understanding all aspects of the internal differentiation of language within a speech community, and even more important for linguistics, the processes of language change. Indeed, it is in these areas where sociolinguistics has had its greatest impact.

To gather data for the study of variation, we note each occurrence of each variant or allophone in a corpus and code it for relevant phonological, morphological and syntactic features of the environment, as well as style, discourse type and any other pertinent situational or extra-linguistic features including the sociodemographic characteristics of the speakers.

The working hypothesis underlying the analysis of the choice of one variant or another of a linguistic variable, which is what we call the *set of possible variants*, is due to LABOV (1969): that the various factors influencing this choice contribute their effects more or less independently of the other factors. We have been formalizing this (SANKOFF, 1975) as follows. Suppose there are two variants. Then

$$\log \frac{P}{1-P} = \mu + a + b + \dots$$

where *P* is the probability that the first variant will be used in a specific context, and *a*, *b*, ... are the effects due to various features or components of this context. μ is an average tendency pertinent to all contexts. The '*a*', for example, refers to the effect of a given feature; it is included in the formula if and only if the corresponding feature is present in the context. This type of formula, with $\log(P/(1-P))$ depending linearly or additively on a number of effects, is well-known in statistics (e.g. COX, 1970; HABERMAN, 1974), for modeling the dependence of probabilities on experimental parameters. Its manipulation tends to be more complicated than ordinary regression or analysis of variance, but presents no fundamental difficulties. JONES (1975) and others have developed programs

to analyze data, i.e. to estimate the factor effects, according to this model and we have also produced a number of programs specifically adapted to large sets of linguistic data, with their multiplicity of contextual features, often subject to co-occurrence constraints.

Note that in the formula, whatever the effects of the factors, as long as they are finite, P will be strictly between zero and one, i.e. the context will be variable and not categorical—speaking in terms of the model, not necessarily of the few observations we may have in this context. We will return to this point later.

The purpose of the data analysis, of course, is to accurately assess and compare the effects of the different features on the occurrence of the variants. With respect to phonological, morphological and syntactic features, this gives rise, as we have mentioned, to linguistic generalizations which are more detailed and refined than those obtainable from a qualitative approach. Thus while it is well known that the // in *il* 'he, it' can be reduced (or absent) in French, careful study reveals that // -loss pervades the system of determiners and pronouns, that it is least stable in impersonal *il* and becomes increasingly stable as personal, plural, and feminine features are carried by the pronoun, that subject clitics are less stable than complements or determiners, and that a consonantal environment is more favorable to // -loss than a vocalic one (G. SANKOFF & CEDERGREN, 1971).

As for extralinguistic features, coding speakers according to their age is a powerful, though not infallible, way of detecting and measuring the progress of linguistic change. Incorporating social class or speaker's sex in the analysis enables us to measure the social differentiation of language and to pinpoint the social origin of change processes. Thus workers drop // more frequently than bourgeois speakers (or obversely, bourgeois speakers insert it more frequently) in all linguistic contexts. Men drop it more frequently than women.

4. The syntactic variable

In turning from phonology to syntax, we encounter a major problem: that of equivalence, i.e. how to ascertain which structures or forms may be considered variants of each other and in which contexts. This was rarely a problem in phonology since invariance of meaning under substitution of allophones is a widely valid criterion, easy to apply, because it depends only on the recognition of two lexical forms as semantically

identical or qualitatively different. Trying to substitute non-equivalent forms leads to discrete changes of meaning—viz. the concept of ‘minimal pair’. This is not the case in syntax. The substitution of different forms often leads to very subtle semantic distinctions, and there is frequently cause for debate as to whether or not there is any change.

Thus the unequivocal basis for phonological equivalence served sociolinguistics as well as it served other approaches to phonological theory, but as in other theories, when notions of syntactic equivalence were required, disagreement and controversy prevailed. This is certainly not because notions of equivalence are peripheral to modern syntactic theory. On the contrary, equivalence, semantic invariance, meaning-preservation, paraphrase and like concepts are basic, the various transformational theories being cases in point.

Early extensions of sociolinguistic analysis to morphosyntactic variation borrowed the notion of the semantic equivalence of transformed and untransformed sentences from generative transformational grammar. Instead of counting two phonological variants in a variety of contexts, studies of syntactic variation counted, under a variety of conditions, the number of sentences in which a given optional transformation had applied versus the number where it had not, even though it could have. Each case was checked to see whether the contrasting variant could be substituted without changing the sense of the sentence.

This approach proved to be unsatisfactory for two kinds of reasons. First, the assumption that transformations are always meaning-preserving is even less defensible for speech in context than in discussion based on intuitions. Second and more important, forms which seem to be equivalent to each other in context, frequently could only be derived by way of very different transformational paths. One could not be considered the transform of the other. Some examples are: different verbal constructions used to convey the same semantic time and aspect: *je ferai / je vais faire*, ‘I will do’ (SANKOFF & THIBAULT, 1978) and different complementizer constructions for a given type of verb complement: *je fais ce qu'il veut / je fais qu'est-ce qu'il veut*, ‘I do what he wants’ (KEMP, 1979).

This latter problem did not occur in phonological variation, incidentally, since phonological variants may generally be analyzed as being generated through the application of rules within the generative phonological framework.

Once the transformational criterion of equivalence is lost, and indeed any criterion based on structural relationships, what is there beside in-

tuition about identical reference which might allow us to postulate equivalence? The quantitative use of natural speech data provides a novel criterion leading to new and far-reaching theoretical consequences.

The clue for the discovery of this criterion existed already in the work on phonological variation. As we have mentioned, coding of tokens for sociodemographic characteristics of speakers leads to the discovery of internal differentiation of the speech community with respect to the variable concerned. There is a weak sort of complementarity of distribution in operation here. In those segments of the speech community where one variant is much used, the other is rarely used, and vice versa. Where one variant is used in moderate amounts, so is the other. Now, looking at this relationship backwards, if we could somehow identify such a pattern of internal differentiation in the community, might this help us to identify the variable involved? We suggest the answer is yes, as long as we have some further grounds, no matter how meagre, for postulating equivalence. That is, we do not have to prove any type of strict semantic invariance. For example, one might expect, and indeed it is empirically the case, that in a given speech community, in a specific speech situation, discourse of a particular type (e.g. narrative) on the same topic will require the expression of completed past action at a certain rate (e.g. once per hundred words), or of restricted relativization at a certain rate. And from our understanding of the discourse, we can decide, with some degree of confidence, when these rather broad functions are being fulfilled. Then, after counting the occurrences of the syntactic structures carrying out these functions, we can test for weak complementarity of distribution. If one form which can fulfill the function is abundant in one segment of the community while another construction is absent or nearly so, if the reverse pattern holds somewhere else in the community, and if the two competing forms are both of moderate abundance elsewhere, then we have weak complementarity and hence equivalence. Of course, this is discourse equivalence or functional equivalence and not necessarily semantic equivalence of syntactic forms—if such a relationship could ever be defined to the satisfaction of all linguists. The boundary between semantics and discourse function is unlikely ever to be a matter of general consensus—there is little agreement even on the details of the semantics/syntax distinction. What is important here is that whatever distinctions the analyst may wish to draw between two forms showing weak complementarity, these distinctions can have little consequence for the speakers, for the communicative uses of language within the community. Consider

for example the case where weakly complementary forms are distributed according to the age of the speaker. Here one form can usually be thought of as replacing the other over time. If they are not syntactic equivalents at some point in time, then what is? This leads, parenthetically, to one often expressed point of view—there are no syntactic variables. But this is clearly a matter of definition. In any case, there is some important type of variation occurring at the syntactic level which has far-reaching implications for the process of linguistic change. For example, using weak complementarity, we were able to confirm (SANKOFF & THIBAULT, 1978) the incipient process of suppletion of *aller* 'to go' in the *passé composé* in the French spoken in Montreal—while one segment of the community uses the expected *je suis allé* 'I went', another use almost exclusively *j'ai été* which, out of context, would be glossed 'I have been'. Weak complementarity shows that they serve identical functions in discourse for different members of the community, a result which would have been inaccessible without interpreting, counting and statistically analyzing thousands of tokens of these forms in natural speech. In another study, we have produced the same type of evidence for equivalence in the use of *avoir* and *être* as auxiliaries in compound tenses for a class of intransitive verbs, despite hypotheses of aspectual distinctions proposed by others on the basis of introspection.

This type of consideration becomes central to a sociolinguistic theory of syntactic change. When discourse equivalents coexist in weak complementarity over a period of time, we may expect this equivalence to become grammaticalized—functional equivalents become syntactic equivalents, with consequent ramifications for the whole syntactic system.

In the pronoun system of Montreal French, LABERGE (1977) studied the almost complete replacement of the first person plural conjugation in French, using the subject clitic *nous* 'we', by the third person singular conjugation using *on*, traditionally glossed 'one'. She showed that this has been at least partly grammaticalized in that non-cliticized co-referents to the subject *on* are generally derivatives of *nous*, as in *on a fait notre mieux* 'one did our best'. Other repercussions are the displacement of *on* from its indefinite usages by *ils* 'they' and *tu* or *vous* 'you' to avoid ambiguity. Thus an important segment of the syntactic apparatus has undergone adjustment due to the ongoing grammaticalization of what was once a stylistic discourse variable.

Based on many other studies of the syntax of Montreal French, it is

our belief that this type of discourse-directed change is at least as important if not more so than the processes of gradual internal grammatical restructuring we find in traditional theories.

5. On waves

Once we have established criteria for the identification of linguistic variables, phonological or syntactic, other questions arise over the analysis of variation data.

There have been two paradigms within which linguistic change and variation have been studied based on natural speech data. One approach which we have discussed breaks down the context of use of a variable into component features and assigns each possible feature a relative weight in favouring one variant over another. These weights are then combined according to a simple model in order to arrive at the overall tendency of the given context. The other approach rejects the decomposition of linguistic context effects into feature effects and prefers to compare entire contexts.

Thus the 'wave theory' (e.g. BAILEY, 1973) envisages linguistic change, i.e. from one variant to another, originating in a highly specific linguistic context in the discourse of a particular segment of the speech community. As time progresses, a wave of change proceeds along both linguistic and extralinguistic dimensions. The original innovators, or their descendants, use the new variant in more and more general contexts, while speakers distant from the centre of innovation start to use it, but only in the original restricted environments. This is illustrated in Figure 1 where the contexts in which the new variant first appeared are to the left and the most conservative contexts to the right. The innovating speakers are at the top, the conservative ones at the bottom. The dotted lines represent the position of the wave 'front' at successive points in time and the heavy line represents its position today.

How does this compare with the other way of looking at change, as represented by the formula

$$\log \frac{P}{1-P} = \mu + a + b ?$$

If the different values of a represent the tendencies of different speakers to use the new variant, and the values of b symbolize the effects due to the various contexts, consider what happens when we rank the speakers

and the contexts in order of these parameter values. As μ increases with time, without any change in the a or b , the values of P increase throughout the diagram, with the highest values in the upper left-hand corner. For a fixed threshold, say $P = 0.5$, if we divide the contexts where the new variant is used more than this proportion of the time from those where

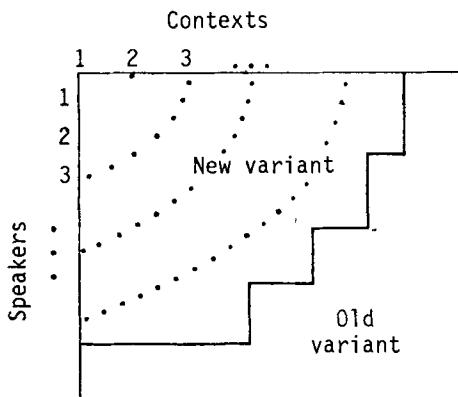


Fig. 1

it is used less, we obtain a wave front just like one of the dotted lines in the diagram. And as time increases, this wave sweeps across the figure from upper left to lower right, just as in the wave theory.

What then is the difference between the two approaches? The main difference is that in strict versions of the wave theory the new variant supersedes the old in cells directly below and to the right of cells already having the new variant, and this replacement must be relatively complete in one cell before it spreads rightward to its neighbour. In other words, $P = 1$ in cells above and to the left of the wave front and $P = 0$ below and to the right. But we already know that these categorical values are not attainable through our formula as long as the parameters are finite. According to the formula, then, all cells will be variable at the same time.

It would seem that an examination of the data would suffice to choose between these two models, but the situation is somewhat more complicated. Data on syntactic variation tends to be very tedious to collect, as one may have to listen to many hours of tape-recordings of natural speech before finding more than a few tokens. And the analytic enthusiasm of linguists knows no bounds, so that the more data they collect the more contexts they will discover. This leads to very sparse data arrays. And

if there are zero or one or even two or three tokens in a cell, it is likely that they will be mostly categorical in appearance, i.e. all tokens in a cell of the new variant or all of the old variant. And it will generally be possible to order speakers and contexts to produce a configuration as in the diagram. In addition, less strict versions of the wave theory will permit a narrow band of variable cells along the wave front as well as a few inconsistent cells in an array, called 'scaling errors'. Thus, choosing among the two theories on the basis of data is no easy matter, and this is reflected in the literature where certain data sets have been analyzed in opposing ways by different authors.

A recent mathematical discovery, however, has changed the way we look at this problem. It is true that the formula will not produce $P = 1$ or $P = 0$ as required by the wave theory as long as μ and the a and the b are all finite. But in analyzing the data to estimate these parameters according to the fundamental criterion of maximum likelihood it is sometimes found that some parameters tend to become infinite. This renders the formula inoperative as it stands, but the natural extension of the maximum likelihood principle leads to well-defined estimates for the cell probabilities as follows: suppose the data can be arranged as in Figure 2 where the 'mixed' blocks do not overlap with respect to rows or columns.

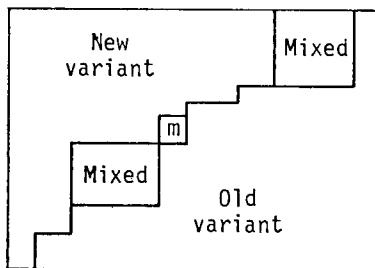


Fig. 2

Then the maximum likelihood estimates of P will be 1 in the upper roughly triangular region and zero in the lower region, i.e. categorical. In the 'mixed' regions P will be between zero and one (variable).

Thus the analysis by the formula and the wave theory analysis converge, with the help of this new theorem (ROUSSEAU & SANKOFF, 1978). Whereas previously the strict form of the wave theory which denies the simultaneous existence of variability in many contexts, and the additive model analysis which postulates simultaneous variability in all contexts, each had their

own methodologies whose results naturally tended to support their respective theories, the integrated mathematical approach now available (SANKOFF & ROUSSEAU, 1979) permits an objective assessment of the relative importance of variability and categoricity in a data set.

6. Conclusion

The two debates we have sketched illustrate the outcome of the socio-linguistic method. They could hardly have been conceived before they arose in the study of quantified natural speech data. Yet they are crucial to understanding important linguistic processes and their formal expression.

References

- BOURDIEU, P., and L. BOLTANSKI, 1975, *Le fétichisme de la langue*, Actes de la Recherche en sciences sociales, vol. 4, pp. 2-31
- BAILEY, C.-J. N., 1973, *Variation and linguistic theory* (Center for Applied Linguistics, Arlington, Virginia)
- COX, D. R., 1970, *The analysis of binary data* (Methuen, London)
- HABERMAN, S. J., 1974, *The analysis of frequency data* (University of Chicago Press)
- JONES, R. H., 1975, *Probability estimation using a multinomial logistic function*, Journal of Statistical Computation and Simulation, vol. 3, pp. 315-329
- KEMP, W., 1979, *L'histoire récente de ce que, qu'est ce que et qu'osque à Montréal: trois variantes en interaction*, in: *Le Français parlé: Etudes sociolinguistiques*, ed. P. Thibault (Linguistic Research Inc., Edmonton)
- LABERGE, S., 1977, *Etude de la variation des pronoms sujets définis et indéfinis dans le français parlé à Montréal*, Ph. D. thesis (Université de Montréal)
- LABOV, W., 1963, *The social motivation of a sound change*, Word, vol. 19, pp. 273-309
- LABOV, W., 1966, *The social stratification of English in New York City* (Center for Applied Linguistics, Washington, D. C.)
- LABOV, W., 1969, *Contraction, deletion, and inherent variability of the English copula*, Language, vol. 45, pp. 715-762
- LABOV, W., 1972a, *Sociolinguistic patterns* (University of Pennsylvania Press, Philadelphia)
- LABOV, W., 1972b, *Language in the inner city* (University of Pennsylvania Press, Philadelphia)
- LABOV, W., A. BOWER, D. HINDLE, E. DAYTON, M. LENNIG, and D. SCHIFFRIN, 1978, *Linguistic change in Philadelphia*, Technical Progress Report on NSF Grant (U. S. Regional Survey, Philadelphia)
- LABOV, W., P. COHEN, C. ROBINS, and J. LEWIS, 1968, *A study of the non-standard English of Negro and Puerto Rican speakers in New York City* (U. S. Regional Survey, Philadelphia)
- LABOV, W., and T. LABOV, 1977, *Learning the syntax of questions*, in: *Recent advances in the psychology of language*, eds. R. Campbell and P. Smith (Plenum Press, New

- York), Also as: *Das Erlernen der Syntax und von Fragen*, Zeitschrift für Literaturwissenschaft und Linguistik, vol. 23/24. Also as: *L'apprentissage de la syntaxe des interrogations*, Langue Française, vol. 34, pp. 52–80
- LABOV, W., M. YAEGER, and R. STEINER, 1972, *A quantitative study of sound change in progress* (U. S. Regional Survey, Philadelphia)
- ROUSSEAU, P., and D. SANKOFF, 1978, *Singularities in the analysis of binomial data*, Biometrika, vol. 65, pp. 603–608
- SANKOFF, D., 1975, VARBRUL version 2 (unpublished program and documentation)
- SANKOFF, D., 1978, *Linguistic variation: Models and methods* (Academic Press, New York)
- SANKOFF, D., and P. ROUSSEAU, 1979, *Categorical contexts and variable rules*, in: Proceedings of the Scandinavian Conference on Syntactic variation, ed. S. Jacobsen (Stockholm)
- SANKOFF, D., and P. THIBAULT, 1978, *Weak complementarity: Tense and aspect in Montreal French*, in: Proceedings of the Conference on Syntactic Change, eds. D. Strong and B. Johns (Ann Arbor, Michigan)
- SANKOFF G., 1980, *The social life of language* (University of Pennsylvania Press, Philadelphia)
- SANKOFF, G., and H. J. CEDERGREN, 1971, *Some results of a sociolinguistic study of Montreal French*, in: *Linguistic diversity in Canadian society*, ed. R. Darnell (Linguistic Research Inc., Edmonton), pp. 61–87
- THIBAULT, P., 1979, *Le Français parlé: Etudes sociolinguistiques* (Linguistic Research Inc., Edmonton)

ON THE METHODOLOGICAL PROBLEMS OF THE HISTORY OF SCIENCE: AN ANALYTICAL APPROACH

MAURICE A. FINOCCHIARO

University of Nevada, Las Vegas, Nevada, U.S.A.

1. Introduction

The methodological problems of the history of science have been discussed by S. R. Mikulinski in a recent issue of the journal *Scientia* (MIKULINSKI, 1975; see also MARKOVA, 1977). He makes a number of interesting points which are worth emphasizing: that scientific biographies are of great importance in the historiography of science (p. 84); that scientists' histories of science have unquestionable significance, even though they by no means exhaust the field (p. 84); that it is wrong to attribute irrationalism to Tomas Kuhn, though one can question many of his theses (p. 86); that the study of the history of science is relatively independent of general history and of philosophy (p. 88); that the historian of science should not only describe scientific development but also "reveal the logic of this development, the regularities underlying the progress of knowledge in a given field" (p. 89); that though Marxists were the first to emphasize the role of social practice in history, "it is completely impossible from a Marxist point of view to infer all complex phenomena in the history of science from the economic conditions, to consider the development of science as determined exclusively by external factors, to ignore the relative independence and activity of consciousness" (p. 92); that the "widespread idea that the Marxist conception of the development of science is externalist is wrong" (p. 95); and that Kuhn's main merit is his emphasis on scientific change "in opposition to positivists whose analysis ususally doesn't proceed further than the investigation of static knowledge" (p. 95). These insights are acceptable, and the discussion

in whose context they occur performs a valuable service insofar as it draws attention to a cluster of problems that need attention.

However, Mikulinski's approach to the definition, systematization, and solution of "the methodological problems of the history of science" is too limited and one-sided. It is too limited insofar as he seems to deal primarily with the 'internalism-externalism' issue, namely with the question of the relation between intellectual and social factors in scientific development; and it is too one-sided in the sense that Mikulinski's own discussion (at the meta-level of the methodology of history of science) is somewhat externalistic, that is, emphasizes social developments in the field and neglects the conceptual analysis of its problems. Thus I believe that his discussion can be supplemented by the one given below, where I emphasize philosophical and logical issues.

The methodological problems of the history of science may be systematized in terms of the notion of a philosophy of the history of science. At the same time this idea helps to conceptualize the various practices in that increasingly popular field which includes the work of philosophically minded historians of science and of historically minded philosophers of science. I am referring to such writers as Joseph Agassi, Paul Feyerabend, Mary Hesse, Thomas Kuhn, Imre Lakatos, Larry Laudan, Ernan McMullin, and Stephen Toulmin.

To begin with, the philosophy of the history of science is the philosophy of two distinct things each of which can be denoted by the term 'history of science'. The first is the discipline history of science, namely what historians of science do, and may be called *historiography of science*. The second is the historical development of science, namely what scientists do and have done in the course of history. These two branches are best called *philosophy of the historiography of science* and *philosophy of the historical development of science*. This distinction is basically the same as the ones that could be made by adapting other terminologies prevalent in the philosophy-of-history literature. For example, adapting Danto's terminology, (DANTO, 1968, Chapter I, esp. p. 1), we could speak of analytical philosophy of the history of science and of substantive philosophy of the history of science, respectively. And adapting Walsh's terminology, (WALSH, 1960, pp. 13-15), we could speak of critical philosophy of the history of science and of speculative philosophy of the history of science. However, I believe my terminology is clearer, less misleading, and less prejudicial.

2. Philosophy of the historiography of science

The philosophy of the historiography of science is simply the philosophy of a particular science or discipline. As such it studies topics like the aim, methods, presuppositions, and logical structure of the historiography of science. To study such matters is no longer a luxury but rather a methodological duty for the philosopher, now that the historiography of science is being institutionalized into a profession. Some philosophers have already felt such a duty; there exist, in fact, at least three works dealing with various aspects of the subject: Agassi's *Towards an historiography of science*, my *History of science as explanation*, and Laudan's *Progress and its problems* (AGASSI, 1963; FINOCCHIARO, 1973; LAUDAN, 1977).

But the philosophy of the historiography of science is not just a matter of duty for the philosopher. It is and ought to be a matter of great interest since one finds in the historiography of science special methodological problems, which either have no counterpart in other fields or else become so magnified and intensified in this field as to provide illuminating and instructive examples.

The two problems I shall discuss here are that of wisdom after the event and that of history-of-science explanation. The latter is the main topic in the above-mentioned *History of science as explanation*. History-of-science explanation may be nominally defined as the explanation of facts and events in the historical development of science, such as a historian of science may be expected to give. Since the concept has been so widely used in the analysis of the structure of both scientific and historical knowledge, one might expect that the analysis of the historiography of science in terms of that concept would be primarily an exercise in the application of the theory of explanation, which has been rather extensively articulated by philosophers. When the application is attempted, although the exercise turns out to be instructive, the results indicate that history-of-science explanation is anomalous.

For example, from the point of view of the so-called *deductive* or *covering law model*, all adequate explanations of events have two properties: (1) they are such that if the explanatory information had been available before the event, that information would have been sufficient to predict it; and (2) they are such that if the explanation is fully stated then the explicans contains at least one universal law or generalization. Using

a rather suggestive terminology to refer to these properties, the model claims that all adequate explanations of events are (1) potentially predictive and (2) law-coverable. Now, in the historiography of science one of the most important class of events which the historian of science may be expected to explain are scientific discoveries. According to the model, then, adequate explanations of scientific discoveries must be potentially predictive and law-coverable. Is that possible?

As regards predictiveness, the question is whether an adequate explanation of a scientific discovery can be such that the explanatory information could have been sufficient to predict the discovery if the information had been available beforehand. Now, to say that the explanatory information would have been sufficient to predict the discovery if the information had been available beforehand means that the occurrence of the discovery would have been inferable from the explanatory information before the discovery took place. But this is to say that the discovery could have been made before it occurred. So according to the model, an explanation of a scientific discovery should show that the discovery could have been made before. And this is an absurd requirement.

As regards law-coverability, the question is whether an adequate explanation of a scientific discovery must be such that the explicans contains universal laws if it is fully stated. This amounts to subsuming the discovery, directly or indirectly, under some law of discovery. But even if the historian were or could be in possession of such laws of discovery, which he is not and perhaps could not be, that would mean that he would be showing that the discovery was not a discovery. Now it may be true in a particular case that what the historian thinks was a discovery, is not really a discovery. But this could not be true in general. In other words, if law-coverable explanations of scientific discoveries were possible, there could be *no* scientific discoveries, no *actual* ones at least, only alleged ones.

Thus the principles of potential predictiveness and of law-coverability turn out to be useful not in historiography of science, but in the philosophy of the historiography of science, the use here being to allow us to put into words and conceptualize the anomaly of the explanation of discoveries.

The problem of wisdom after the event, which is the main one discussed in Agassi's *Towards an historiography of science*, is the following. In the investigation of a given history-of-science episode, what is the proper use of the scientific knowledge which was brought about during that episode or which came into being thereafter? For example, if scientific

knowledge tells us that nothing of scientific value was accomplished during a certain episode, would there be any point for the historian of science to study it? Would it even be likely that he would perceive the episode at all as falling within his area of concern? And when he starts ignoring many episodes which only later can be seen to be scientifically worthless, is he not going to get a biased view of the historical development of science? Is he not going to necessarily see that development as straightforwardly progressive and cumulative?

It might seem that this is basically a practical problem for the historian of science and has no philosophical significance. I would not want to deny for a moment that it is a practical problem; the extent to which it has been and remains so has been well documented in Agassi's book. But I also believe that wisdom after the event does have philosophical importance, that it tells us something very interesting and peculiar about the nature of the historiography of science. In fact, the possibility that is emerging is that there is a field, namely the study of the historical development of scientific knowledge, where real and genuine bias is introduced by knowledge itself, where a certain kind of ignorance is better than knowledge.

I might put it this way. The old problem of whether ignorance is bliss dealt with the desirability of knowledge and the extent of the conflict between knowledge and happiness. Our reflections on the historiography of science add a new dimension to this problem, a dimension that goes deeper into the nature of knowledge. For it may be that ignorance is bliss in an epistemological sense, that *some* ignorance is desirable for the sake of *some* knowledge, that knowledge of certain things makes impossible knowledge of other things. If this is true, its philosophical importance could hardly be overestimated.

I am not concerned here with trying to establish that fact but only in suggesting that it may very well be true and in pointing out that the seriousness of its possibility is best seen in the case of scientific knowledge and the historical knowledge of its development.

3. Philosophy of the historical development of science

Let us now go on to consider the philosophy of the history of science in the sense of the philosophy of the historical development of science. This is, first of all, an approach to the philosophy of science. In this approach one philosophizes about past science as well as about con-

temporary science, instead of philosophizing merely about the latter, as nonhistorical philosophers of science do. Thus the philosophy of the history of science in this sense is a branch of philosophy that can include the nonhistorical approaches to the philosophy of science, though the latter cannot include it. In other words, philosophy of the history of science so conceived is philosophy of science-in-its-various-historical-stages. There is no special problem with this idea.

One can also think of the philosophy of the historical development of science as the philosophy of the *history* of something. The problem then arises whether such a philosophy is possible. For it would seem that, though there is much to study and investigate about the development of something, there is nothing to philosophize about. What can there be to do in our case above and beyond what the historian of science does?

I believe various answers to this question are possible. First one can study the development of science in a philosophically explicit manner. One way to do this would involve the more or less conscious practical application, in the actual study of concrete historical episodes, of ideas in the philosophy of the historiography of science. Another way would involve the reflective extraction, after an actual historical study, of ideas of the philosophy of the historiography of science. Both of these activities would admittedly be mixtures of historiography of science and of philosophy of the historiography of science, but I do not think it would be improper to categorize such mixed inquiries as philosophy of the historical development of science. Laudan's *Progress and its problems* is perhaps the best example of philosophy of history of science in this sense.

Second one could study the fundamental features and patterns of the historical development of science. If this study were conducted at a sufficiently high level of generality, and if the developmental outlines were sufficiently broad, then such a study would be philosophical. The reason for this is that by such studies one would acquire perspective, and though the search for intellectual perspective is not always a philosophical activity, when the matter is sufficiently important, then that search becomes philosophy. To have an over-all view of the historical development of science is nowadays extremely important given its cultural supremacy in the West and its material supremacy in the whole world.

I believe this point can be made clearer with the help of an analogy. Consider the fact that philosophy includes the study of such things as the relationship between mind and body and the existence (or non-existence) of God. The reason why the study of such things is regarded

as part of philosophy is really historical. That is, there were times when, and there exist systems of thought where, a view concerning those matters is relevant to practically every aspect of human experience. The situation is analogous regarding the over-all development of science. There may be a question as to whether the relevant systems of thought exist, but certainly there can be no question that in our own time the view one has about the fundamental features of the development of science is relevant to practically every aspect of contemporary human experience. Hence, if we do not yet have a system of thought in which such a relevance is demonstrated and illustrated, it is our task as philosophers to accept the challenge.

It is in this category of philosophy of the history of science in the sense of theory of the over-all pattern of the historical development of science that I would put such works as KUHN's (1970) *Structure of scientific revolutions* and much of TOULMIN's (1972) *Human understanding*. What we have here is basically the same thing that the historian of science does, but at a greater level of generality.

Next, there is a third kind of investigation whose philosophical character could hardly be questioned, but whose legitimacy is in doubt. This would be the attempt to formulate the concepts in terms of which the development of science is to be understood. If the concepts in question were empirical concepts, there would be nothing to distinguish this activity from either the historiography of science or the philosophy of the historical development of science in one of the previously mentioned senses. But if the concepts could be *a priori*, or somehow nonempirical, then we would have a philosophy in the noncontroversial sense of an analysis of the concept 'development of science'. So the question becomes, Do we have or can we formulate a nonempirical concept of the development of science? The nonempirical must here be understood not as absolutely *a priori* but in the sense in which the concept of explanation is nonempirical; one does not just dream up the concept but rather studies it by examining aspects of human experience relevant to it.

In our present case some relevant experiences turn out to be the discussions carried out by historically minded philosophers of science and by philosophically minded historians of science at the 1965 International Colloquium in the Philosophy of Science. An in-depth, philosophical analysis of its proceedings as reported in Lakatos and Musgrave's *Criticism and the growth of knowledge*, (LAKATOS & MUSGRAVE, 1970; cf. FINOCCHIARO, 1972/73), reveals that the historical development of science

is to be conceived in terms of three distinct but interrelated concepts: scientific change, scientific progress, and scientific rationality. These three concepts can be derived essentially from the kinds of questions that it is conceivable to ask about a given episode, and that fact is what gives those concepts their philosophical character. The concept of scientific change relates to the question of how and why the episode took place. Scientific progress pertains to the question of the sense in which the episode was a change for the better. Scientific rationality relates to the question of the nature of the reasons held by the historical agents from whose behavior the episode resulted.

With these distinctions in mind we may say that the construction of a general theory of scientific progress (so distinguished), and the construction of a general theory of scientific rationality (so distinguished) represent two other species of the philosophy of the historical development of science. From the present point of view the construction of a general theory of scientific change would basically correspond to the construction of what I earlier called a theory of the over-all pattern of the historical development of science. Or to be more exact, there would be such a correspondence if one pursued the latter with the conceptual purism being suggested here. If one is not too dogmatic about such purism, then one could allow, at least at a metalevel, a conceptually mixed theory of the over-all development of science, which would then be distinct from a general theory of scientific change *per se*. One could make this allowance as long as the conceptual mixture was not a chaos, that is as long as the logical interrelationships among our trilogy of concepts was sound in the actual practice. For in this case the problem would be no greater than that of having three different topics somewhat intermingled in the same book. If the interrelationships among change, progress, and rationality were unsound in the actual practice, that would mean that one would be supporting conclusions about change with historical evidence about rationality or progress, conclusions about progress with evidence about change or rationality, and conclusions about rationality with evidence about change or progress. The accomplishment would then be nil.¹

And this raises the question of the exact nature and interconnections among this triad of concepts, which may be regarded as one of the topics

¹ Laudan's theory of scientific growth (in Part I of LAUDAN, 1977) is an example of such an innocuous mixture of theory of rationality and theory of progress, whereas *Criticism and the growth of knowledge* is an example of such an unsound mixture of reflections on change, on progress, and on rationality.

covered by the philosophy of the history of science in the sense of the analysis of the concept 'historical development of science'. To give a brief statement of my own view about this, I might begin by asking the question. Why bother about a general theory of scientific change? Of course, one may want to bother for purely intrinsic reasons, namely to gain understanding. If, however, one could put a theory of scientific change to an extrinsic use, then one would have greater motivation to undertake the task. I believe that some fundamental features of scientific change would have interesting consequences. For example, suppose scientific change proceeds in a particular direction; then this direction could be used to define progress. That is, a theory of change could ground a theory of progress. The argument would then be that science has changed for the better because it has normally proceeded in such and such a direction, e.g. there has been an increase of understanding. Could a theory of progress ground a theory of rationality? That is, could one argue that a good scientific reason is to want to bring about progress, with progress specified in a sufficiently detailed manner so that the reason would acquire some content? It seems to me we could. Then we would have a theory of rationality grounded on one of progress, grounded in turn on one of change.

If book titles are any indication, I would say that the distinctions I have made here among change, progress, and rationality are implicit in TOULMIN's trilogy on *Human understanding* (1972, p. ii). The first volume deals with what he calls "the collective use and evolution of concepts"; this probably corresponds to a theory of scientific change. The second volume, not yet published, will deal with "the individual grasp and development of concepts"; here I would expect to find a theory of scientific rationality. The third volume, also unpublished, will deal with "the rational adequacy and appraisal of concepts"; here I would expect to find a theory of scientific progress. What Toulmin's view is of the interconnections, and whether he would agree with the one sketched above, is not within the scope of this inquiry to determine.

In summary, the varieties of philosophy of the history of science that have so far emerged are: (1) philosophy of the historiography of science; (2) philosophy of past as well as contemporary science; (3) historiography of science in synthesis with a philosophy of the historiography of science; (4) theory of the over-all patterns of the historical development of science; (5) conceptually distinct general theory of scientific change; (6) conceptually distinct general theory of scientific progress; (7) conceptually distinct

general theory of scientific rationality; and (8) theory of the nonempirical aspects of the concept 'historical development of science', or more specifically, theory of the nature and interconnections among the distinct concepts of scientific change, scientific progress, and scientific rationality.

4. Logic of actual scientific reasoning

Finally, there is a field of philosophical problems, challenges, and opportunities deriving from the history of science which deserves special attention. Though this enterprise is little known and little practiced, it would be difficult to overestimate the mutual benefits that historians of science and philosophers could derive from it. The activity deserves a special term, and so I shall refer to it as *history-of-science logical criticism*. This concept may be defined as the critical understanding of reasoning and arguments actually put forth by individual scientists in the course of the history of science. By *critical understanding* here I mean the analysis, reconstruction, and evaluation of such actual arguments, conducted in such a way as to avoid uncritical elucidation and misunderstanding criticism.

Such an inquiry is a species of philosophy of the history of science in the sense of logic of actual scientific reasoning. That is, one is studying those aspects of the historical development of science that are amenable from a logical point of view. And those are the aspects for the study of which a philosopher has potentially a superior preparation since the theory and practice of reasoning constitutes such a central part of philosophical inquiry.

Before discussing several examples of history-of-science logical criticism, it will be useful to examine the concept from other points of view. First, it is an engrossing and delightful type of inquiry which one may occasionally undertake for its own sake. It is the analogue for the world of thoughts and ideas of what literary criticism is in the world of images and expressions. Of course, the development of science is not the only place where one can find arguments that can become the subject matter of logical criticism; the history of philosophy is obviously another such domain, and the histories of jurisprudence and of historiography are also full of reasoning. But there is no need here to get into fruitless and invidious comparisons.

Second, history-of-science logical criticism could be regarded as the much sought after way of combining history of science and philosophy

of science. At any rate it would constitute a genuine synthesis of the two fields. For what I am suggesting here is that the combination be carried out as follows: the philosopher takes the development of science as his subject matter, and then he exploits the skill in which he excels because of his training, namely the critical analysis of arguments. Of course, to succeed he also needs some acquaintance with the details of the historical knowledge situation. But the critical understanding of actual scientific arguments would be a *rational reconstruction* of the history of science in the best sense of the term.

Third, history-of-science logical criticism could be regarded as an approach to the philosophy of science, if philosophy is equated to logic and logic to the study of reasoning, and if we equate science to actual, historical science as contrasted to the 'science' in someone's theory of science. Philosophy of science would thus be the study of reasoning in actual, historical science.

Fourth, the critical understanding of actual scientific reasoning would constitute an approach to the theory of scientific rationality, in the above-mentioned sense according to which one distinguishes rationality, progress, and change in science.

Fifth, history-of-science logical criticism would be *applied logic* insofar as it would be the application of the logician's principles, concepts, and techniques to actual scientific reasoning. The principles being applied could come either from general logical theory, or from symbolic logic, or from inductive logic. It is obvious, however, that such an application might turn out to be more than a mere application; it could become a testing of the theoretical principles. If one was more interested in the formulation and testing of logical principles by reference to actual scientific reasoning, then one would have applied logic in the sense of an approach to logic, an approach that can also be called, and has been called, *practical logic*, or *informal logic*. Though scientific practice is not the only domain that could be studied in this approach to logic, it offers many advantages such as the explicitness of its records, and epistemological respectability.

Such is history-of-science logical criticism at the level of philosophical theory. Let us now ask how it performs in practice. Unfortunately, as I have already mentioned, there is little work done in the field. The underlying cause of this condition is that the enterprise requires both logical sophistication and acquaintance with the historical knowledge situation, and historians of science typically lack the former, philosophers the latter.

Moreover, when philosophers do apply themselves to the study of history-of-science arguments, they usually do so to make philosophical points, philosophical in the sense of epistemological or methodological. They do not aim at making either purely logical and purely historical points. Such has been the case for three arguments I shall discuss below, namely Newton's Third Rule of Philosophizing, Galileo's rejection of causal investigation, and Galileo's space-proportionality argument. There is nothing wrong with studying such arguments to make epistemological or methodological points, other than that such an activity is well known and widely practiced, and hence not the one presently under discussion. The point which I wish to make, and which I will illustrate below, is that it is possible to analyze such arguments to make points that are purely historical and/or purely logical (as distinct from epistemological and methodological). And since this is so, there exists a nearly virgin field of opportunities for philosophers.

Let us now examine some examples. The first pertains to Newton's Third Rule of Reasoning in Philosophy. In the standard English edition of the *Principia* the rule is stated as follows: "The qualities of bodies, which admit neither intensification nor remission of degrees, and which are found to belong to all bodies within the reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever" (NEWTON, 1934). A two-page commentary follows this statement. The logical analysis of this commentary, carried out in the spirit of the enterprise under consideration (cf. FINOCCHIARO, 1974a), shows that there are really two distinct rules in Newton's mind. One rule would state that a quality is an essential quality of a certain class of objects only if it makes no sense to attribute it to a greater or lesser degree and it belongs to all of those objects that have been observed. The other rule would state that a quality is to be regarded as present throughout the universe, or as present universally, or (simply) as universal, if it belongs to all observed objects. Since the above-quoted English statement of the rule contains no apparent ambiguity, one is led to question of the accuracy of the translation. This reveals that, corresponding to the unambiguous English phrase 'universal qualities of all bodies whatsoever', the original Latin has a phrase which has a sufficiently subtle ambiguity as to render comprehensible Newton's equivocation in his commentary. One is also led to formulate two as yet untested historical hypotheses: (1) that the Latin in that commentary contains evidence for the ambiguity detected from a logical analysis of the English, or else further evidence of the deficiency

of the standard English translation; and (2) that drafts of the two different rules exist among Newton's manuscripts relating to this so-called 'third rule of reasoning in philosophy'.

It should be noted that this rule is a typical example of something that has previously attracted the attention of philosophers. But they have been concerned primarily about its role in Newton's derivation of gravitation. Whereas the analysis summarized above was intended to provide a service to historians of science by reaching some historical conclusions and suggesting further historical working hypotheses.

The next example is Galileo's refutation of space-proportionality. (GALILEI, 1953, p. 160, and 1890-1909, vol. VIII, pp. 203-204.) Space-proportionality is the false idea that the speed of freely falling bodies is proportional to the space traversed from rest. Physics tells us that time-proportionality is true instead, namely their speed is proportional to the time elapsed since the beginning of the fall. Galileo's argument has stirred many controversies, because it appears to be fallacious. The controversies were recently rekindled when Stillman Drake discovered that the major modern translations of the Italian passage are demonstrably inaccurate and then suggested that this inaccuracy spread and infected the usual reconstructions of the argument. (DRAKE, 1970a, 1970b.) The results that emerge when the argument is analyzed in the spirit of history-of-science logical criticism are the following. (See FINOCCHIARO, 1972, 1973.)

First, one traditional interpretation alleged that Galileo was invalidly using the so-called *mean-speed theorem*, and since this theorem was widely known in the Middle Ages as the Merton Rule, this interpretation was used as part of the evidence suggesting a Medieval origin of Galileo's ideas. The logical analysis of the passage shows this interpretation to be untenable and also to have been untenable before Drake's discovery of the erroneous translation, on which the interpretation was likely to be based. Thus some of the work done by Drake's historical scholarship could have been accomplished by logical analysis.

Drake's own interpretation, which I believe also to be inaccurate, is untenable in part because he attributed to Galileo a question-begging stratagem which any philosopher would have been sufficiently acquainted with so as to avoid making the attribution. Galileo is trying to refute space-proportionality on the ground that it would imply instantaneous motion. What is really needed is a justification of this implication and what appears erroneous is precisely that inference. The following argument is, however, formally valid: if space-proportionality is true,

then free fall motion is instantaneous; but free fall motion is not instantaneous; therefore, space proportionality is false. If we attribute this argument to Galileo, as I believe Drake does, then Galileo would be committing no formal logical error. But the problem with his argument would not disappear, for it could now be rephrased as the question of whether the first premise of the argument is true. In other words, any argument whose validity is dubious can be rewritten as an argument whose validity is assured but whose soundness remains dubious. This is so well known to a logician that either he would not attribute this question-begging stratagem to Galileo, or else he would want to expand the argument so as to include some other propositions from which the crucial conditional premise might have been derived.

The third result that emerges from the present example, is that when Galileo's argument is reconstructed in the most accurate and plausible possible manner, it turns out to be an awe-inspiring instance of the fallacy of equivocation. I say this, because the history-of-science logical critic will feel that from Galileo's argument he can learn what fallacies of equivocation really are. He will not have to rely any longer on examples such as: American buffalos are practically extinct; this animal is an American buffalo; therefore, this animal is practically extinct (COPI, 1978). The reconstructed Galilean argument is the following:

- (1) If the speed of a falling body increases as the space, that means that the speeds with which it passes a given space are doubles of the speeds with which it passes the first half of that space.
- (2) Now, if the speeds with which a falling body passes a given space are doubles of the speeds with which it passes the first half of the space, then that means that its speeds are as the spaces passed or to be passed.
- (3) But, when the speeds are as the spaces passed or to be passed, such spaces are passed in equal times.
- (4) Therefore, if the speed of a falling body increases as the space, then the various spaces of different lengths over which it passes are passed in equal times.
- (5) But spaces of different lengths are passed in different times by a falling body.
- (6) Therefore, it is false that the speed of a falling body increases as the space.

The trouble lies in the fact that the consequent of (2) does not mean the same as the antecedent of (3). The former means that the speeds (of the

falling body) are as the total spaces passed to acquire those speeds; the latter means that the speeds are as the spaces passed at those speeds.

The next example regards the famous passage in *Two new sciences* where Galileo is alleged to have renounced the investigation of causes. (GALILEI, 1974.) Here I believe that logical criticism can solve the long standing historical problem of what exactly Galileo meant. In this case it turns out that the context in which the passage occurs must be examined in order to understand the passage properly. This context is what the logically sensitive reader is led to examine in order to ascertain the logical point of the passage; its logical point is precisely what most discussions of the passage neglect, even the discussions by philosophers, who in the present case are too quick to make epistemological or methodological points.

When the passage is examined in the spirit of history-of-science criticism (see FINOCCHIARO, 1975/76) it emerges that the renunciation of causes is part of an argument designed to support the comprehensibility (as distinct from the truth) of the principle of time-proportionality, namely that falling bodies acquire equal increments of speed in equal times. Galileo is arguing that the principle is comprehensible, because and to the extent that (among other reasons) it can be used together with other comprehensible ideas without it being accepted as true. These other ideas are elements of the impetus theory, which Galileo speculates could be combined with time-proportionality to explain why falling bodies accelerate at all. It is after this speculative passage that comes the one where Galileo asserts he is not really concerned with the cause of the acceleration of falling bodies. That remark has the logical tone of a qualification and hence has the function of showing that Galileo does not accept the speculative explanation of acceleration previously discussed but is instead trying to illustrate how the idea of time-proportionality, whose comprehensibility was in question in the context, could be combined with other ideas, whose comprehensibility was not in question, but whose truth could very well be questioned. And that would presumably justify the comprehensibility of time-proportionality.

The next example concerns the argument from falling bodies which was perhaps the single most famous physical argument against the motion of the earth. It is examined and criticized by Galileo in the "Second Day" of the *Dialogue concerning the two chief world systems* (GALILEI, 1953, pp. 139–141, and 1890–1909, vol. VII, pp. 164–167). By studying the relevant passages the history-of-science logical critic can find what is perhaps the single best instance of *petitio principii*, or at least the best

source-material for the formulation of the concept. A reconstruction of the falling bodies argument is the following (for the details, see FINOCCHIARO, 1974b): The earth does not move because bodies thrown vertically upwards fall vertically back to the same place from which they were thrown, and this could not happen if the earth were moving; for if the earth were moving, then the place of ejection would move along with it while the projectile was in the air going up and down, and if that place moved, then the body would fall some distance away from it. Galileo's criticism of this argument is that it begs the question insofar as the Aristotelians would have to support their premise that bodies fall vertically by means of the following argument: Bodies fall vertically because they are *seen* to fall vertically and apparent vertical fall implies actual vertical fall. And they would have to justify this last premise as follows: apparent vertical fall implies actual vertical fall, because *if* the earth does not move, then apparent vertical fall does imply actual vertical fall, and the earth does not in fact move. Now, this last premise is identical with the final conclusion of their falling bodies argument.

My last example of a logical opportunity in the study of the history of science concerns the concept of *ad hominem* argument. Once again one finds excellent source-material in Galileo who uses the concept on several occasions (see FINOCCHIARO, 1974c). This time, however, his concept is definitely different from ours. In fact, he uses the term in such a way that an *ad hominem* argument is not a fallacy, not even an alleged fallacy, but rather an argument which uses premises accepted by an opponent, but not necessarily by the arguer, in order to derive conclusions not acceptable to the opponent. From this one obviously learns something about the history of the term 'ad hominem argument': and one can learn something about the history of the concept, if one discovers some kind of internal development. Moreover, one could learn something about the nature of argument in general, if C. L. Hamblin's theory is correct that an essential feature of the concept of argument is its being "dialectical" in an Aristotelian sense which has some correspondence with the old meaning of the term 'ad hominem argument' (HAMBLIN, 1970, especially Chapt. 7). One instance of such *ad hominem* argument occurs in Galileo's *The Assayer* (DRAKE & O'MALLEY, 1960, pp. 29-30; cf. FINOCCHIARO, 1974c): It is not true that the upper part of the earth's atmosphere is carried around by the revolution of the lunar orb since this orb must be attributed the perfect shape of a sphere and air is not swept along by its mere contact with the smooth surface of this revolving sphere; that

this is so can be shown from the experiment of a candle placed inside a concave circular vessel with a smooth internal surface; the flame of the candle does not move at all no matter how fast we revolve the vessel. Galileo's opponent here is someone who held a different view about the nature of comets. In the argument just stated Galileo is deriving a conclusion which denies one of his opponent's views from his opponent's belief in the smoothness of the lunar sphere. On the other hand, Galileo does not even accept the existence of the lunar sphere. (Other examples can be found in FINOCCHIARO, 1977, and FINOCCHIARO, 1979 and 1980.)

To conclude, in this paper I have tried to systematize the methodological problems of the history of science accordance with what may be called an analytical approach, by contrast, for example, to S. R. Mikulinski's discussion of the same problems. I have done so in terms of the idea of a philosophy of the history of science, which subsumes two main concepts, that of the philosophy of the historiography of science and that of the philosophy of the historical development of science. The latter in turn splits into three main branches, depending on whether the emphasis is on science, on philosophy, or on development. That is, if one studies the development of the *philosophical aspects of science*, we have what was called above philosophy of past (and present) science, and this corresponds to what is ordinarily called 'history of the philosophy of science'; if one studies scientific development *per se* by means of a *philosophical approach*, then we get, for example, either a synthesis of historiography of science with philosophy of the historiography of science, or an analysis of the concept 'historical development of science', or an informal logic of actual scientific reasoning; if one studies the *philosophical aspects of the development of science*, then we have the theory of the general patterns of the evolution of science.

References

- AGASSI, J., 1963, *Towards an historiography of science* (History and theory, Beiheft 2) (Mouton, Hague)
- COPI, I. M., 1978, *Introduction to logic*, 5th ed. (MacMillan, New York)
- DANTO, A. C., 1968, *Analytical philosophy of history* (Cambridge University Press, Cambridge)
- DRAKE, S., 1970a, *Galileo studies* (University of Michigan Press, Ann Arbor), pp. 229–239
- DRAKE, S., 1970b, *Uniform acceleration, space, and time*, British Journal for the History of Science, vol. 5, pp. 28–43
- DRAKE, S., and C. D. O'MALLEY (eds. and trans.), 1960, *The controversy of the comets of 1618* (University of Pennsylvania Press, Philadelphia)

- FINOCCHIARO, A., 1972, *Vires acquirit eundo: The passage where Galileo renounces space-acceleration and causal investigation*, Physis, vol. 14, pp. 125–145
- FINOCCHIARO, A., 1972/73, *Review-essay on “Criticism and the growth of knowledge”*, eds. I. Lakatos and A. Musgrave, Studies in History and Philosophy of Science, vol. 3, pp. 357–372
- FINOCCHIARO, A., 1973a, *Galileo’s space-proportionality argument: A role for logic in historiography*, Physis, vol. 15, pp. 65–72
- FINOCCHIARO, A., 1973b, *History of science as explanation* (Wayne State University Press, Detroit)
- FINOCCHIARO, A., 1974a, *Newton’s third rule of philosophizing: A role for logic in historiography*, Isis, vol. 65, pp. 66–73
- FINOCCHIARO, A., 1974b, *Galileo as a logician*, Physis, vol. 16, pp. 129–148
- FINOCCHIARO, A., 1974c, *The concept of ad hominem argument in Galileo and Locke*, The Philosophical Forum, vol. 5, pp. 394–404
- FINOCCHIARO, A., 1975/76, *Cause, explanation, and understanding in science: Galileo’s case*, The Review of Metaphysics, vol. 29, pp. 117–128
- FINOCCHIARO, A., 1977, *Logic and rhetoric in Lavoisier’s Sealed Note: Toward a rhetoric of science*, Philosophy and Rhetoric, vol. 10, pp. 111–122
- FINOCCHIARO, A., 1979, *The logical structure of Galileo’s Dialogue*, Logique et analyse, vol. 22, pp. 159–80
- FINOCCHIARO, A., 1980, *Galileo and the art of reasoning* (Reidel, Dordrecht)
- GALILEI, Galileo, 1890–1909, *Opere*, edited by A. Favaro, this and later editions (Barbera, Florence)
- GALILEI, Galileo, 1953, *Dialogue concerning the two chief world systems*, translated by S. Drake, this edition and later reprints (University of California Press, Berkeley)
- GALILEI, Galileo, 1974, *Two new sciences*, translated by S. Drake (University of Wisconsin Press, Madison)
- HAMBLIN, C. L., 1970, *Fallacies* (Methuen, London)
- KUHN, T. S., 1962, *The structure of scientific revolutions*, 1st ed. 1962, 2nd ed. enlarged 1970 (University of Chicago Press, Chicago)
- LAKATOS, I., and A. MUSGRAVE, eds., 1970, *Criticism and the growth of knowledge* (Cambridge University Press, Cambridge)
- LAUDAN, L., 1977, *Progress and its problems: Towards a theory of scientific growth* (University of California Press, Berkeley)
- MARKOVA, L. A., 1977, *Difficulties in the historiography of science*, in: Historical and philosophical dimensions of logic, methodology and philosophy of science, eds. R. E. Butts and J. Hintikka (Reidel, Dordrecht), pp. 21–30
- MIKULINSKI, S. R., 1975, *The methodological problems of the history of science*, Scientia, vol. 110, pp. 83–97
- NEWTON, Isaac, 1934, *Mathematical principles of natural philosophy* (A. Motte’s 1729 translation revised by F. Cajori, 1934.) (University of California Press, Berkeley)
- TOULMIN, S. E., 1972, *Human understanding*, vol. I (Princeton University Press, Princeton)
- WALSH, W. H., 1960, *Philosophy of history* (Harper, New York)

ON THE CHANGE OF THE INTERRELATIONS BETWEEN SCIENCE AND EPISTEMOLOGY OF SCIENCE IN THE PROCESS OF THEIR HISTORICAL DEVELOPMENT

V. LEKTORSKY

U.S.S.R. Academy of Sciences, Moscow, U.S.S.R.

1. In the investigation of the epistemological problems of science which constitutes the bulk of philosophy of science it has often been assumed that,

- (a) The determining epistemic characteristics of science, the canons of rationality, are eternal and invariable;
- (b) Scientific activity and the reflection on this activity—the epistemological analysis of science—belong to two different domains, the first one being entirely independent of the second.

Epistemology of science attempts to expose the norms of activity which are accepted in science and at the same time, revealing the epistemic status of these norms, it deals with philosophical problems which are not directly linked with a scientist's activity. The process of generating scientific knowledge is independent of the scientist's awareness of the norms governing it. This process belongs to what K. Popper calls "the third world", the world of objective mind and objective knowledge.

2. However, both these assumptions seem rather questionable.

One of the specific features of scientific activity is that it presupposes a more or less developed process of critical reflection on its own premisses. Reflection thus is not something external to science, but incorporated into it, "built into" its body. And it is this feature which distinguishes science from myth, on the one hand, and from technology, on the other.

Not only does reflection ensure the normal functioning of science, it also serves as a means of its re-building, which often involves the changing of the epistemic norms and canons of rationality. In other words, ideas about the acceptability of some or other ontological assumptions, about

the correctness of some or other modes of reasoning, about the link between theoretical constructions and empirical data, etc., serve as a means of developing and remaking the content of scientific knowledge.

The situation in mathematics in 19th century offers here a good example. The activity in the field of critical analysis—from the reformation of mathematical analysis to arithmetization of mathematics and its foundation on the basis of set theory—was not merely an activity aimed at revealing the logical structure of mathematical knowledge. It was also an activity of constructing (and at the same time, remaking) the very building of mathematical science.

In this connection it should be noted that this activity necessitated the tackling of a number of epistemological problems bearing on the understanding of the nature of mathematical knowledge.

Epistemological, philosophical, methodological, and metaphysical ideas of Galileo, Descartes, Newton were indispensable for the constitution of classical science. And it is due to the analysis of epistemological problems carried out by Einstein, Bohr, Heisenberg, that the theory of relativity and the quantum theory have emerged.

3. In works on philosophy of science one may come across the opposition between the scientists' image of their activity and what they are really doing. Bridgeman, as is known, advised those who deal with epistemology and methodology of science not to rely on scientists' words but to scrutinize their activity. No doubt, there is a certain discrepancy between the procedures really fulfilled in science, the epistemic norms used in it, and the way the scientist understands them.

Here we do not only mean to say that scientists' consciousness only partially reflects the procedures which they are making use of. One should also take into account the fact that in some cases the reflection can produce a quite distorted image of science.

Thus, the discrepancy between Newton's actual scientific activity and his explicit methodological prescriptions has become an almost trite example. In this respect it is interesting to note that it was Newton's epistemological foundation for the results obtained—namely his claims to having "deduced" the principal laws of mechanics directly from the empirical phenomena—that provided the condition for a special theoretical status to be given to these laws in the system of science.

This, in its turn, led to the transformation of science and to the creation of the basis for classical mechanics—the solid basis upon which it stood

for about two hundred years. Thus not only is Newton's epistemological and methodological reflection at variance with his own scientific practice, it is also presupposed by the latter as an indispensable component.

So anyone concerned with the problems of epistemology of science should not only take into consideration what scientists are actually doing but also the way they interpret their own activity, what they "say" about it, for these "sayings" very often prove to be an important component of the activity itself.

Moreover, what scientists "actually do" in science is by no means obvious by itself. There is no other way to reveal the actual meaning of scientific activity than through the reflection on science.

4. We sometimes come across an opposition between the epistemological self-consciousness of scientists and the work on the problems of epistemology of science which is carried out by professional philosophers. The scientists' reflection—as internal to science—is considered as an indispensable element of the development of science. As far as the epistemological reflection of professional philosophers is concerned, it is looked upon as something purely "external" to science and aimed at tackling problems which lie beyond the scientist's concern.

To be sure, there is a difference between a scientist's reflection and reflection in the framework of epistemology of science as a special branch of philosophy. The scientists' reflection is connected with problems of epistemology and methodology in so far as these problems are important for the understanding of the nature and of the ways of development of one or another branch of science. As to the epistemology of science as an intellectual activity performed by professional philosophers, it deals with problems of epistemic status of different structural elements of science in a very broad context; as a rule, it correlates scientific knowledge with specific features of science in general; it reveals logical links between different layers of scientific knowledge; it dwells on certain ontological problems, etc. In other words, the epistemology of science as a specific branch of knowledge aims at creating an image of science as a whole and thus, at solving all the problems bearing on this task.

Naturally, a scientist as a specialist in a particular branch of science may be quite indifferent to all these problems.

It should be noted, however, that the scientist's reflection and the professional philosopher's reflection on science are not totally independent of one other. A scientist may find himself facing the problem of scientific

knowledge in the broad sense—and in this case his reflection may be transformed into professional philosophical reflection. This, in its turn, may give an impetus to the analysis of purely scientific problems on a new methodological basis.

The opposition between the tasks of science and the tasks of its philosophical interpretation is of a rather recent origin and is not characteristic of classical science and classical philosophy. For Descartes and Leibniz science and epistemology of science are equally important and interconnected domains. But even after philosophy and scientific activity were brought apart, they were still, for quite a long time, often thought of as belonging to one and the same system of knowledge, where all the components were mutually determined.

Classical epistemology was, on the one hand, a critical analysis of some primary assumptions of classical science, and, on the other hand, its theoretical foundation and intellectual sanction, a way of incorporating science into the system of knowledge in general, a way to interpret its ontological status.

Classical epistemology of science is the intellectual prerequisite of classical science. Here we do not mean one or another version of this epistemology (and these versions, as is well known, were essentially different and quite often opposed to each other) but epistemology of science as a specific type of intellectual activity. All this by no means renders invalid our previous statement that in a number of ways classical epistemology was a distorted image of classical science.

Later there arose certain periods in the history of epistemology of science when the unity of scientific activity and its image created by the philosophy of science, was not perceived. Here we mean the type of analysis of science which was introduced by the school of logical positivism.

5. One of the characteristic features of classical epistemology is that though it claimed to reveal the universal structure of science, it was in fact, nothing more than an analysis of and a philosophical foundation for a certain type of science, classical mechanics being regarded as a paradigm of scientific construction. What this epistemology actually did was to make an absolute of a certain type of science and of a corresponding intellectual attitude by attributing them to the mechanism of a cognizing activity treated either on the level of individual consciousness or of super-individual cognizant mechanisms.

These versions of epistemology of science were, in fact, prescriptive.

This was so even when they contained criticism of some premises of classical science.

It is also noteworthy that science was recognized as supreme form of cognizing reality and at the same time science in general was identified with the historically determined type of science which was embodied in the mechanistically oriented sciences.

This accounts for the specific way of treating and discussing problems of general epistemology, which was characteristic of the philosophy of the XVII–XVIII centuries. Such problems as the interrelation of “primary” and “secondary” qualities, the logical status of statements about necessary connections and other questions which caused a lot of debate at that time presupposed the image of science, knowledge and reality imposed by classical science.

In the course of the scientific revolutions of the 20th century the image of scientific knowledge and knowledge in general has undergone certain changes. It has become a widespread notion that there is no absolute correlation between one or another system of knowledge and the corresponding domain of objective reality, that the degree of this correlation is historically determined. There is further evidence that objective reality is dialectical and contradictory in itself, that there exists a number of layers which are not reducible to one other.

It follows from what has been said above that there is a possibility of the co-existence of several theories of different types which are realizations of different intellectual attitudes and different modes of relations between theoretical and empirical statements and different types of scientific explanation.

It also means that epistemic norms may be and inevitably are different and that all attempts to find some theories which would serve as an “absolute basis” for theoretical knowledge in general are useless.

Hence the necessity of re-building epistemology of science in a certain way in order to study objective correlations occurring in the process of development of scientific knowledge and to reveal and to systematize different types of attitudes possible in science. In other words, epistemology of science acquires a definite dialectical dimension.

(a) The importance for epistemology of science of investigating the history of science by no means leads to the former transformation into the latter. Epistemology of science turns to the history of science in so far as it is necessary in order to reveal the intellectual attitudes in science and to analyze their historical development.

Epistemology of science as a branch of philosophy studies science in the context of knowledge as a whole and touches upon certain ontological problems.

(b) Epistemology of science should take into consideration the fact that it is, like science, historically determined. So, not only history of science, but history of epistemology of science itself, too, represents a very important object of investigation for epistemology of science. An examination of history of epistemology of science enables us to observe various approaches in the philosophical analysis of science and to analyze their strong and weak points. An inquiry into the history of epistemology of science may prove fruitful in the sense that it may reveal some ideas which are of vital importance in the present-day situation.

Thus, for instance, some of the ideas bearing on the understanding of scientific knowledge and knowledge in general which were developed in German classical philosophy turn out to be of more value for epistemology of science now than the mode of analysis suggested by logical positivism.

The doctrine of "*a priori* foundations of a pure science of Nature" of Kantian philosophy may serve as a good example. Kantian apriorism is by no means acceptable, this doctrine deriving from the dogmatization and canonization by Kant of the contemporary mechanistically-oriented science. However, Kant was the first in epistemology of science to direct attention to the specific role of primary principles underlying any large theoretical research program. He has shown that these principles differ in the way in which they are revealed and justified both from mere empirical generalizations as well as from particular hypothetico-deductive systems.

Hegel was the first in the history of epistemology to show the universal dialectical mechanism of the development of knowledge as a process of an infinite revealing of the incompleteness of its foundations and of the creative process of rebuilding these foundations. According to Hegel, reflection of knowledge on itself at every stage of its development is an incomplete reflection in that it presupposes the presence of movements of consciousness, which have not "undergone" reflection, which are performed "behind the back of consciousness", so to speak.

Being aware of something in this or that form does not mean knowing it as Hegel says. With Hegel, knowledge is a universal dialectical process in the course of which both subject and object undergo transformation. The subject is not an individual object, something primarily self-contained,

but perpetual movement, becoming, development, the overcoming of all established boundaries, and the constant positing of new ones. It cannot be thought of without the object which it is knowing and changing. The object, too, is transformed in the process of the development of consciousness, i.e., it changes in the historical process of the cognitive activity.

The Hegelian idealistic conception, which in the long run reduces knowledge to reflection and makes the object of knowledge dependent on self-consciousness, cannot be accepted. However, his conception of dialectical interrelation in the development of knowledge between reflection and the contents of consciousness which have not "undergone" reflection seems to bear directly on the actual problems of epistemology of science.

It would be stressed that the rational moments of Hegel's understanding of science and knowledge have been critically modified and assimilated by the Marxist epistemology of science.

ON THE ORIGIN AND SUBSEQUENT APPLICATIONS OF THE CONCEPT OF THE LINDENBAUM ALGEBRA

STANISŁAW J. SURMA

University of Sokoto, Nigeria, Jagiellonian University of Cracow, Poland

I. BIOGRAPHY OF ADOLF LINDENBAUM

Very little seems to be known about the life and work of the author of the well-known *Lindenbaum algebras* and *Lindenbaum extension lemma*. To begin with, the exact date of his birth remains obscure. Small wonder, the archives one might have gathered this information from were destroyed during the Second World War, and so we have to rely solely upon the oral tradition or other similarly indirect and thus not always reliable sources. According to some of those sources, Adolf Lindenbaum was born around 1904 (cf. FRANKEL, 1967, and FRANKEL and BAR-HILLEL, 1958) most probably in Warsaw. However, some other conceivable dates have also been mentioned, for instance, 1905, cf. MESCHKOWSKI (1964). Lindenbaum's early childhood remains completely obscure, as do his years spent in secondary school. He entered Warsaw University soon after the end of the First World War. There he studied Mathematics and Logic mainly under the guidance of Professors W. Sierpiński, J. Łukasiewicz and S. Leśniewski. Again, almost nothing is positively known about his studies. Even the titles of his Ph. D. and D. Sc. (Docent) theses have not been identified, let alone their contents. One thing we know for sure is that he became Doctor of Philosophy not later than 1927 and Doctor of Science (Docent) before the academic year 1936/37; his name was entered as Dr. Adolf Lindenbaum in the 1927 list of members of the Polish Mathematical Society and he was mentioned as Doc. Dr. Adolf Lindenbaum in the *Calendar of Warsaw University for the year 1936/37*. It is also known that Professor W. Sierpiński was one of his D. Sc. thesis supervisors.

In 1936, Lindenbaum married Miss Janina Hossiasson born in 1899. The bride was thus 37 years old. Was the bridegroom really five to six years her junior?

A. Lindenbaum took an active part in the academic life of both the Polish and international scientific communities. He presented some of his results at meetings of such bodies as the Philosophy and Mathematics Students Association (see papers L1, L2, L7 in the subsequent Bibliography of Adolf Lindenbaum); at the three consecutive Congresses of the Polish Mathematicians (papers L6, L10, L18, L19, L34, L35); at the First Congress of the Slav Mathematicians (paper L9); at the Logical Section of the Warsaw Philosophical Society (papers L30, L33); at the Polish Logical Society (papers L38, L39); and at the International Congress of Scientific Philosophy held in Paris, 1935 (paper L32).

A. Lindenbaum was elected Secretary of the Polish Logical Society at its foundation meeting in 1936 at the same time when J. Łukasiewicz was elected Chairman and A. Tarski Vice-Chairman. Lindenbaum served as one of two Co-Chairmen along with E. Szpiłrajn, editor of Section I: *Foundations of Mathematics, Set Theory and Theory of Real Functions*, IIIrd Congress of Polish Mathematicians, Warsaw, 1937. In 1938 he became Treasurer of the Warsaw Branch of the Polish Mathematical Society. He was also Assistant of the Logical Seminar, which was conducted by Professor J. Łukasiewicz, at the Warsaw University from 1926 to the outbreak of the Second World War. Beginning with the academic year 1936/37, A. Lindenbaum delivered lecture courses on the following topics:

- (i) On new investigations into the foundations of mathematics and the mathematical foundations of other disciplines (1936/37 academic year);
- (ii) On superposition of functions (1936/37 and 1937/38 academic years);
- (iii) Chosen topics from metrology and from the theory of functions (1938/39 academic year).

A. Lindenbaum as well as his wife met his death at the hands of the Nazis. The tragedy took place most probably in the summer of 1941. However, there is no general agreement as to its place. Some sources mention the Warsaw ghetto, cf. GROMSKA (1948). Some others indicate the town of Białystok, cf. EPITAH (1945a) or the ghetto in Nova Vileyka, a town near Vilno, cf. EPITAH (1945b).

Lindenbaum's work falls into the following areas:

- (i) set theory and its foundations;
- (ii) logic, among others, the Lindenbaum extension lemma, the Lindenbaum algebras and a number of results on Łukasiewicz's many-valued logics;
- (iii) methodology, especially definability theory, and the concept of simplicity.

His scientific activities resulted in 26 articles and 14 abstracts of lectures, 40 published items in total. Some of his results were published during his lifetime by others, e.g. the Lindenbaum extension lemma was published by A. Tarski, cf. TARSKI (1930a), while the method of the Lindenbaum algebras was first presented by J. C. C. McKinsey (1941).

II. BIBLIOGRAPHY OF ADOLF LINDENBAUM

Below is Lindenbaum's complete bibliography or so it seems in light of the current state of research into the work and life of A. Lindenbaum.

- L1. *O podstawach matematyki* (On the foundations of mathematics, Polish), Ruch Filozoficzny, vol. 9, 1925, p. 117
- L2. *O równoważności układów aksjomatycznych logistyki B. Russella i D. Hilberta* (On the equivalence of the axiom systems for the logics of B. Russell and D. Hilbert, Polish), Ruch Filozoficzny, vol. 9, 1925, p. 117
- L3. *Communication sur les recherches de la théorie des ensembles* (With A. Tarski), Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie, Classe III, vol. 19, 1926, pp. 299–330
- L4. *Contribution à l'étude de l'espace métrique. I*, Fundamenta Mathematicae, vol. 8, 1926, pp. 209–222
- L5. *Sur l'arithmétique des types ordinaux*, Annales de la Société Polonaise de Mathématique, vol. 5, 1926, pp. 103–104
- L6. *O matematycznych metodach badania nad teorią dedukcji* (On mathematical methods of investigations into the theory of deduction, Polish), Ruch Filozoficzny, vol. 10, 1926/27, p. 205
- L7. *O pojęciu nieskończoności* (On the concept of infinity, Polish), Ruch Filozoficzny, vol. 10, 1926/27, p. 209
- L8. *Sur l'indépendance des notions primitives dans les systèmes mathématiques* (With A. Tarski.), Annales de la Société Polonaise de Mathématique, vol. 5, 1927, pp. 111–113
- L9. *Z podstaw teorii grup* (On the foundations of the theory of groups, Polish), Ruch Filozoficzny, vol. 11, 1928/29, p. 201

- L10. *Méthodes mathématiques dans les recherches sur le système de la théorie de deduction*, Księga Pamiątkowa Igo Polskiego Zjazdu Matematycznego we Lwowie w 1927 r., Supplement to Annales de la Société Polonaise de Mathématique, 1929, p. 36
- L11. *Remarques sur une question de la méthode axiomatique*, Fundamenta Mathematicae, vol. 15, 1930, pp. 313–321
- L12. *Sur les opérations d'addition et de multiplication dans les classes des ensembles* (With A. Koźniewski), Fundamenta Mathematicae, vol. 15, 1930, pp. 342–355
- L13. *Bemerkungen zu den vorhergehenden "Bemerkungen" des Herrn J. v. Neumann*, Fundamenta Mathematicae, vol. 17, 1931, pp. 335–336
- L14. *Sur un ensemble linéaire extrêmement non homogène par rapport aux transformations continues et sur le nombre des invariants de ces transformations*, Annales de la Société Polonaise de Mathématique, vol. 10, 1931, pp. 113–114
- L15. *Le projection comme transformation continue le plus général*, Annales de la Société Polonaise de Mathématique, vol. 10, 1931, pp. 116–117
- L16. *Sur les figures convexes*, Annales de la Société Polonaise de Mathématique, vol. 10, 1931, pp. 117–118
- L17. *Sur les constructions non effectives dans l'arithmétique élémentaire*, Annales de la Société Polonaise de Mathématique, vol. 10, 1931, pp. 118–119
- L18. *Formalizacja elementarnego rachunku* (Formalization of the elementary calculus, Polish), Annales de la Société Polonaise de Mathématique, vol. 10, 1931, p. 140
- L19. *Badania nad właściwościami metrycznymi mnogości punktowych* (Investigations into the metric properties of point sets, Polish), Annales de la Société Polonaise de Mathématique, vol. 10, 1931, p. 141
- L20. *Sur les ensembles dans lesquels toutes les équations d'une famille donne ont un nombre de solution fixé d'avant*, Fundamenta Mathematicae, vol. 20, 1933, pp. 1–29
- L21. *Sur les ensembles localement dénombrables dans l'espace métrique*, Fundamenta Mathematicae, vol. 21, 1933, pp. 99–106
- L22. *Sur les superpositions des fonctions représentables analytiquement*, Fundamenta Mathematicae, vol. 23, 1934, pp. 15–37
- L23. *Sur le "problème fondamental" du jeu d'échecs*, Annales de la Société Polonaise de Mathématique, vol. 13, 1934, pp. 124–125
- L24. *Sur les superpositions de fonctions représentables analytiquement*, Annales de la Société Polonaise de Mathématique, vol. 13, 1934, p. 125
- L25. *Sur le nombre des invariants des familles de transformations arbitraires*, Annales de la Société Polonaise de Mathématique, vol. 13, 1934, p. 131
- L26. *Remarques sur la groupe des permutations de l'ensemble des nombres entiers*, Annales de la Société Polonaise de Mathématique, vol. 13, 1934, p. 131
- L27. *Sur les relations contenues dans les relations ordinaires*, Annales de la Société Polonaise de Mathématique, vol. 13, 1934, p. 132
- L28. *Z teorii uporządkowania wielokrotnego* (On the theory of the multiple ordering, Polish), Wiadomości Matematyczne, vol. 37, 1934, pp. 1–35
- L29. *Über die Beschränktheit der Ausdrücksmittel deductiver Theorien*, (With A. Tarski) Ergebnisse eines mathematischen Kolloquiums, vol. 7, 1934/35, pp. 15–22
- L30. *O pewnych kwestiach metodologicznych związanych z podstawami geometrii. Część 1 i 2* (One some methodological questions connected with the foundations of geometry. Parts 1 and 2, Polish), Ruch Filozoficzny, vol. 13, 1935, p. 476

- L31. *Sur le nombre des invariants des familles de transformations arbitraires.* II, Annales de la Société Polonaise de Mathématique, vol. 15, 1936, p. 185
- L32. *Sur la simplicité formelle des notions,* Actes du Congrès International de Philosophie Scientifique, VII Logique. (Actualités Scientifiques et Industrielles, vol. 394) (Paris), 1936, pp. 29–38
- L33. *O zagadnieniach związanych z pewnym kryterium pojęć* (On problems connected with a criterion for the simplicity of concepts, Polish), Ruch Filozoficzny, vol. 13, 1937, p. 149
- L34. *Numerotage des types logiques,* Annales de la Société Polonaise de Mathématique, vol. 16, 1937, p. 191
- L35. *Sur l'équivalence de deux figures par décomposition en nombre fini de parties respectivement congruentes,* Annales de la Société Polonaise de Mathématique, vol. 16, 1937, p. 197
- L36. *Sur l'indépendance de l'axiome du choix,* Annales de la Société Polonaise de Mathématique, vol. 16, 1937, p. 217
- L37. *Sur les bases des familles de fonctions,* Annales de la Société Polonaise de Mathématique, vol. 17, 1938, pp. 124–126
- L38. *O prostocie formalnej* (On formal simplicity, Polish), Ruch Filozoficzny, vol. 14, 1938, p. 151
- L39. *O pracach i projektach międzynarodowej komisji ujednolicenia symboliki logicznej* (On the work and projects of the international committee for the unification of logical symbolism, Polish), Ruch Filozoficzny, vol. 14, 1938, 151
- L40. *Über die Unabhängigkeit des Auswahlaxioms und einiger seiner Folgerungen* (With A. Mostowski), Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie, Classe III, vol. 31, 1938, pp. 27–32

The picture drawn above is by no means complete as Lindenbaum's biography can hardly be documented. Therefore, one has to be aware of a certain danger of oversimplification. However, even the scarce information available seems to be worth presenting as, unfortunately, there is little hope of enriching our current knowledge of A. Lindenbaum's life (cf. also, SURMA, 1973; and SZUMAKOWICZ, 1977).

III. LINDENBAUM ALGEBRAS

The scope of the present paper is limited to discussion of the origins and more significant advances in the method of Lindenbaum algebras. The Lindenbaum algebras and the Lindenbaum extension lemma, by far the most important of Lindenbaum's achievements, play a fundamental role in contemporary model theory and logical semantics.

1. The origins of the method

The method of Lindenbaum algebras was developed in the academic year 1926/27, though its full appreciation dates as recently as the forties in the case of the propositional calculus and the late forties and early fifties as far as the predicate calculus is concerned.

At the foundation of the Lindenbaum method lies an unorthodox idea of Adolf Lindenbaum for building up a logical matrix of the considered propositional calculus out of the same language, i.e. out of the same formulas and connectives which make up the calculus itself. This idea was first made known in the academic year 1926/27 to Professor J. Łukasiewicz's Seminar on Mathematical Logic held at Warsaw University. At that time Lindenbaum applied his idea to the proof of the following theorem:

THEOREM 1. For any set X of zero order propositions closed under substitution of propositional variables and under detachment there exists an at most denumerable normal matrix M such that X coincides with the contents of M .

Note that the theorem remains true if we omit simultaneously the condition that X is closed under detachment and the condition that M is normal. The theorem is one of the most fundamental theorems in the theory of logical matrices. It was published without proof by ŁUKASIEWICZ and TARSKI in their joint paper in 1930.

The idea of treating the sets of propositions of a formalized language as logical matrices was again presented by Lindenbaum in his address to the First Congress of Polish Mathematicians held in Lvov in September, 1927, but it was not included in the summary of the addresses in the Congress Proceedings (cf. L10).

Thus the Lindenbaum method has never been published by its author nor by any of his contemporaries during his lifetime, and it has survived solely through the oral tradition of the Warsaw School of Mathematics. However, it should be emphasized that in the thirties the Lindenbaum method was not appreciated and recognized as a tool of constructing logical matrices even though some similar ideas were brought up at the time. For instance, in 1935 A. Tarski introduced an algebra of propositional formulas with operations corresponding to the propositional connectives. Apart from a different treatment of the identity relation in this algebra it coincided with the Lindenbaum construction, cf. SURMA (1973). A close

relationship with the Lindenbaum method is also evident in the way the proof of completeness of the propositional calculus was presented by M. Wajsberg in 1936 (cf. WAJSBERG, 1937).

The Lindenbaum method came to be known outside Poland thanks to Tarski, one of the most active participants of Łukasiewicz's Seminar in the twenties and thirties. It was published for the first time in 1941 by J. C. C. McKinsey, Tarski's close associate, who referred to it as "unpublished method of Lindenbaum" (cf. MC KINSEY, 1941, p. 222) "explained to me by Professor Tarski" (cf. MC KINSEY, p. 222, footnote 4). McKinsey gave an outline of the proof of only a particular case of Theorem 1.

2. Łoś's contribution

In 1949 J. Łoś published a monograph (Łoś, 1949), an extensive and systematic treatise on the Lindenbaum method which gave rise to the general theory of the Lindenbaum algebras and its applications to propositional calculi. The monograph described a method of constructing two kinds of matrices made up of propositional formulas of a given propositional calculus. The matrices were labelled by Łoś the *m-th Lindenbaum matrices of meaningful expressions*, and the *m-th Lindenbaum matrices of proofs*. The latter matrices were the reduction of the former by special congruences called the *interexchangeability relations* by Łoś. In Łoś (1949) a detailed proof of Theorem 1 was provided.

In detail, the construction is as follows: Let S be the least set of propositional formulas which includes all the propositional variables, p_1, p_2, \dots , and is closed under one binary connective Φ . (For the sake of simplicity, we consider only one connective.) By S_m , where $m \leq n_0$, we denote the set of formulas in S containing no variables other than p_1, p_2, \dots, p_m . Thus, $S_m \subseteq S_{m+1}$ and $S_{n_0} = S$. Let $M = \langle W, V, \varphi \rangle$ be a matrix where $V \subseteq W \neq \emptyset$ and φ is a binary operation on W . M is said to be *normal* if $a \in V$ and $b \in W - V$ entail $\varphi(a, b) \in W - V$. Let $E(M)$ stand for the contents of M , i.e. the set of all formulas obtaining an element of V as a value under any replacement of Φ by φ and of the variables by the elements of W .

$M(X, S_m) = \langle W, V, \varphi \rangle$ is said to be the *m-th Lindenbaum matrix of meaningful expressions determined by $X \subseteq S$ and S_m* iff X is closed under

substitution, $W = S_m$, $V = X \cap S_m$ and $\varphi(A, B) = \Phi(A, B)$ for any $A, B \in S_m$. The following equality holds:

$$(1) \quad E(M(X, S_m)) \cap S_m = X \cap S_m.$$

Hence

$$(2) \quad E(M(X, S)) = X.$$

The formulas A and B are said to be *interexchangeable on the basis of* $X \subseteq S$, in symbols, $A \sim_x B$, iff X is closed under substitution and $C(A) \in X$ is equivalent to $C(B) \in X$ for any $C \in S$. Note that the relation \sim_x or simply \sim is a congruence in S .

$M(X, S_m)/\sim = \langle W, V, \varphi \rangle$ is said to be the *m-th Lindenbaum matrix* of proofs determined by $X \subseteq S$ and S_m iff X is closed under substitution, $W = S_m/\sim$, $V = (X \cap S_m)/\sim$ and $\varphi([A], [B]) = [\Phi(A, B)]$ for any $A, B \in S_m$. The following equality holds:

$$(3) \quad E(M(X, S_m)/\sim) = E(M(X, S_m)).$$

3. Generalizations

The original Lindenbaum method was applied to the propositional calculi. In the late forties and the early fifties many generalizations of Lindenbaum's construction were developed and subsequently used in the first order logic, among others, to prove the following Completeness Theorem due to GöDEL (1930):

THEOREM 2. *Every consistent set of first order sentences has a model verifying it.*

In the proofs of Theorem 2 usually some modification of the following well-known theorem are used:

THEOREM 3. *Every consistent set of first order sentences has a maximal consistent extension.*

Theorem 3 is the well-known *Lindenbaum extension lemma* formulated and proven by Lindenbaum. Again, it was not published by its author. It was only in 1930 that A. Tarski quoted it without proof in his paper (TARSKI, 1930a).

In this paper we will discuss three independent generalizations of the Lindenbaum method due, respectively, to J. Łoś (1949 and 1955), to L. HENKIN (1949), and to H. RASIOWA and R. SIKORSKI (1950).

Łoś's approach. In Łoś (1949), the so-called *algebraic Lindenbaum matrices of proofs* were described and subsequently used in proving Theorem 2 for sets of open first order sentences. Let T denote the least set containing all individual variables z_1, z_2, \dots and closed under functional symbols, in this case, under one binary functional symbol F . Let S_T stand for the least set of sentences containing propositional variables p_1, p_2, \dots , predicates, in this case, one binary predicate $R(t_1, t_2)$, where $t_1, t_2 \in T$, and closed under one binary connective Φ . Obviously, $S \subseteq S_T$. By T_n , where $n \leq \aleph_0$, we denote the set of terms containing no variables other than z_1, z_2, \dots, z_n . Obviously, $T_n \subseteq T$. Let $S_{m,n}$ be the set of sentences containing no variables other than p_1, p_2, \dots, p_m and z_1, z_2, \dots, z_n . Note that $S_{\aleph_0, \aleph_0} = S_T$.

A sequence $\mathfrak{M} = \langle M, \mu \rangle$ is said to be an *algebraic matrix* iff $M = \langle W, V, \varphi \rangle$ is a logical matrix and $\mu = \langle U, f, r \rangle$ is a relational structure such that

- (i) U is a non-empty domain for the individual variables and disjoint from W ,
- (ii) f is a binary operation on U which corresponds to the function symbol F , and
- (iii) r is a binary relation on U , i.e. a function from $U \times U$ to W , which corresponds to the predicate symbol R .

$E(\mathfrak{M})$ stands for the contents of \mathfrak{M} , i.e. the set of sentences true in \mathfrak{M} .

We modify the previously accepted definition of the interexchangeability relation by letting A and B be simultaneously either in S_T or in T . This relation, denote it by \approx , is a congruence in S_T and in T , respectively. $\mathfrak{M}(X, S_{m,n}, T_n)/\approx$ is said to be the (m, n) -th *algebraic Lindenbaum matrix of proofs* determined by the sets $X \subseteq S_T$, $S_{m,n}$ and T_n iff

- (i) X is closed under substitution,
- (ii) X and $S_{m,n}$ determine the m -th Lindenbaum matrix of proofs $M(X, S_{m,n})/\approx$,
- (iii) $f([t_1], [t_2]) = [F(t_1, t_2)]$, and
- (iv) $r([t_1], [t_2])$ holds in U iff $R(t_1, t_2) \in X$, where $t_1, t_2 \in T_n$.

The following equality holds:

$$E(\mathfrak{M}(X, S_{m,n}, T_n)/\approx) \cap S_{m,n} = X \cap S_{m,n}.$$

In Łoś (1949), Theorem 2 was proved for sets of open first order sentences. The author made use of the following lemmas:

LEMMA L1. *Any set of open first order sentences having the property P extends to a maximal set with the same property.*

LEMMA L2. *If X is a maximal set of open first order sentences having the property P, then there exist numbers m and n such that*

$$X = E(\mathfrak{M}(X, S_{m,n}, T_n)/\approx).$$

Here a set of sentences X is said to have the *property P* iff

- (i) X is closed under substitution and detachment, and
- (ii) $X \cap S$ coincides with the set of all tautologies in the two-valued logic.

By the well-known Skolem procedure of eliminating quantifiers, the above method can be applied also to sets of closed first order sentences, which was done in another Łoś paper (Łoś, 1955). Note that a similar way of proving Theorem 2 can be found in E. W. BETH (1951).

Henkin's approach. Another generalization of the Lindenbaum method was presented in L. HENKIN (1949). Henkin first described a denumerable sequence of inductions using in fact Theorem 3. Namely, let S_0 be the set of closed first order sentences without individual constants. The set can be ordered e.g. in the form: $S_0 = \{A_{0,1}, A_{0,2}, \dots\}$. Let $X \subseteq S_0$ be consistent and define

$$(i) \quad Y_{0,0} = X,$$

$$Y_{0,j+1} = \begin{cases} Y_{0,j} \cup \{A_{0,j+1}\} & \text{if } Y_{0,j} \cup \{A_{0,j+1}\} \text{ is consistent,} \\ Y_{0,j}, & \text{otherwise,} \end{cases}$$

$$Y_0 = \bigcup \{Y_{0,j}: j = 0, 1, \dots\}.$$

It can easily be verified that $Y_0 \subseteq S$ and that Y_0 is a consistent and complete set of sentences closed under detachment. Put $Y_0 = \{A_{0,1}, A_{0,2}, \dots\}$ and let $\{t_{1,1}, t_{1,2}, \dots\}$ be a sequence of individual constants not occurring in the elements of Y_0 . Denote by S_1 the set of all sentences of the language extended by $\{t_{1,1}, t_{1,2}, \dots\}$. Define X_1 as follows:

$$(ii) \quad X_{1,0} = Y_0,$$

$$X_{1,j+1} = \begin{cases} X_{1,j} \cup \{B_{1,j}(x/t_{1,j})\} & \text{if there exists a formula} \\ & B_{1,j}(x) \text{ with one free vari-} \\ & \text{able } x \text{ such that } A_{0,j} \\ & = (Ex)B_{1,j}(x), \\ X_{1,j} & \text{otherwise,} \end{cases}$$

$$X_1 = \bigcup \{X_{1,j}: j = 0, 1, \dots\}.$$

Put $S_1 = \{A_{1,1}, A_{1,2}, \dots\}$ and define Y_1 as follows:

$$(iii) \quad Y_{1,0} = X_1,$$

$$Y_{1,j+1} = \begin{cases} Y_{1,j} \cup \{A_{1,j+1}\} & \text{if } Y_{1,j} \cup \{A_{1,j+1}\} \text{ is consistent,} \\ Y_{1,j} & \text{otherwise,} \end{cases}$$

$$Y_1 = \bigcup \{Y_{1,j}: j = 0, 1, \dots\}.$$

It is easy to verify that Y_1 is a consistent and complete set closed under detachment.

Now, assume inductively that $Y_i = \{A_{i,1}, A_{i,2}, \dots\} \subseteq S_i$ where S_i results from S_{i-1} by adding $\{t_{i,j}: j = 1, 2, \dots\}$ as a sequence of new individual constants not occurring in the elements of Y_i and suppose that Y_i is a consistent and complete set closed under detachment. With the help of $\{t_{i+1,j}: j = 1, 2, \dots\}$ we form the set S_{i+1} and we define X_{i+1} as follows:

$$(iv) \quad X_{i+1,0} = Y_i,$$

$$X_{i+1,j+1} = \begin{cases} X_{i+1,j} \cup \{B_{i+1}(x/t_{i+1,j})\} & \text{if there exists a formula } B_{i+1,j}(x) \text{ with one free variable } x \\ & \text{such that } A_{i,j} = (Ex)B_{i+1,j}(x), \\ X_{i+1,j} & \text{otherwise,} \end{cases}$$

$$X_{i+1} = \bigcup \{X_{i+1,j}: j = 0, 1, \dots\}.$$

Put $S_{i+1} = \{A_{i+1,1}, A_{i+1,2}, \dots\}$ and define Y_{i+1} as follows:

$$(v) \quad Y_{i+1,0} = X_{i+1},$$

$$Y_{i+1,j+1} = \begin{cases} Y_{i+1,j} \cup \{A_{i+1,j+1}\} & \text{if } Y_{i+1,j} \cup \{A_{i+1,j+1}\} \text{ is consistent,} \\ Y_{i+1,j} & \text{otherwise,} \end{cases}$$

$$Y_{i+1} = \bigcup \{Y_{i+1,j}: j = 0, 1, \dots\}.$$

Obviously, $Y_{i+1} \subseteq S_{i+1}$ and Y_{i+1} is a consistent and complete set closed under detachment. Finally, we put

$$(vi) \quad S_\omega = \bigcup \{S_i: i = 0, 1, \dots\},$$

$$Y_\omega = \bigcup \{Y_i: i = 0, 1, \dots\},$$

$$T_{\omega,\omega} = \{t_{i,j}: i, j = 1, 2, \dots\}.$$

The following lemma holds:

LEMMA H1. Y_ω is a consistent and complete set closed under detachment and such that for any formula $A(x)$ with one free variable x if $(Ex)A(x) \in Y_\omega$, then there exists $t \in T_{\omega,\omega}$ such that $A(x/t) \in Y_\omega$.

The above construction is comparatively simple, because we have assumed that

- (i) in the elements of S_0 no individual constants occur,
- (ii) we add as new symbols individual constants and not arbitrarily complex terms, and
- (iii) the index j is an integer.

Obviously, all these assumptions can be abandoned. In particular, one can put that $j < \gamma_0$, where γ_0 is an arbitrary transfinite ordinal, to obtain T_{ω, γ_0} instead of $T_{\omega, \omega}$. In the last case the Axiom of Choice must be used.

In fact, Henkin described a model which was essentially the algebraic Lindenbaum matrix $\mathfrak{M}(Y_\omega, S_\omega, T_{\omega, \omega})$ though he did not make any explicit references of Lindenbaum himself. Henkin showed that the following lemma holds:

LEMMA H2. $Y_\omega = E(\mathfrak{M}(Y_\omega, S_\omega, T_{\omega, \omega}))$.

This completes the proof of Theorem 2 by Henkin's method since $X \subseteq Y_\omega$.

The above method of Henkin was originally applied to the classical first order logic. Later on, in 1950, cf. HENKIN (1950), it was extended also to some of the non-classical first order logics. Henkin's method was simplified and modified by many authors. We present below the simplification due to G. HASENJÄGER (1953).

Let S_0 be the set of all closed first order sentences and let $X \subseteq S_0$. Assume that X is consistent and consider the set $T_\omega = \{t_1, t_2, \dots\}$ of individual constants. Denote by S_1 the set of sentences constructed with the help of T_ω . We denote by X^+ the set resulting from X by adding to it the logical axioms expressed in the enriched language involving the new constants. Obviously, $X^+ \subseteq S_1$ and it is easy to verify that X^+ is consistent. Let $\{A_k(x_{i,k}): k = 1, 2, \dots\}$ be the set of all formulas with one free variable $x_{i,k}$. Choose a sequence $\{t_{j,1}, t_{j,2}, \dots\} \subseteq T_\omega$ such that $t_{j,k}$ occurs in none of the formulas $A_1(x_{i,1}), A_2(x_{i,2}), \dots, A_k(x_{i,k})$, for $k = 1, 2, \dots$, and such that $t_{j,k}$ is different from each of $t_{j,1}, t_{j,2}, \dots, t_{j,k-1}$. Denote by B_k the sentence:

$$(Ex_{i,k}) A_k(x_{i,k}) \rightarrow A_k(x_{i,k}/t_{j,k})$$

and put

$$\begin{aligned} X_{1,0} &= X^+, \\ \text{(i)} \quad X_{1,i+1} &= X_{1,i} \cup \{B_{i+1}\}, \\ X_1 &= \bigcup \{X_{1,i}: i = 0, 1, \dots\}. \end{aligned}$$

It is easy to verify that $X_1 \subseteq S_1$ and that X_1 is a consistent set closed under detachment. Finally, put

$$\begin{aligned} Y_{1,0} &= X_1, \\ Y_{1,i+1} &= \begin{cases} Y_{1,i} \cup \{A_{i+1}\} & \text{if } Y_{1,i} \cup \{A_{i+1}\} \text{ is consistent,} \\ Y_{1,i} & \text{otherwise,} \end{cases} \\ Y_1 &= \bigcup \{Y_{1,i} : i = 0, 1, \dots\}. \end{aligned}$$

Obviously, $Y_1 \subseteq S_1$ and there holds a lemma resulting from Lemma H1 by replacing Y_ω and $T_{\omega,\omega}$ by Y_1 and T_ω , respectively.

Note that variants of Henkin–Hasenjäger’s method were discussed also in GRZEGORCZYK (1969) as well as in REICHBACH (1955), SURMA (1968 and 1969).

Rasiowa–Sikorski’s approach. The paper by H. RASIOWA and R. SIKORSKI (1950) concerning the proof of Theorem 2 for sets of closed first order sentences, was published in 1950. A detailed exposition of the authors’ method was also contained in their monograph, RASIOWA and SIKORSKI (1968). The method can be described as follows.

\mathfrak{M} is said to be a *quantifier algebraic Lindenbaum model determined by* $X \subseteq S_1$ and T_ω iff $\mathfrak{M} = \langle M, \mu, \cup \rangle$ where

- (i) M is the two-valued matrix,
- (ii) μ is a relational structure as described above, and
- (iii) \cup is a function from T_ω to the set of all mappings of the set S_1/\approx into itself satisfying the condition:

$\cup \{[A(x/t)] : t \in T_\omega\} = [(Ex)A(x)]$ for any formula $A(x)$ with one free variable x .

In the above definition the existential quantifier is thus conceived of as an operation of infinite meet in a (complete) Boolean algebra. Of course, this idea is not completely original. We can find it, for instance, in MAUTNER (1946) and in MOSTOWSKI (1948).

In proving Theorem 2, Rasiowa and Sikorski made use of two auxiliary lemmas. The first lemma is a modification of the well-known Boolean prime ideal theorem, formulated and proved independently by several authors for different purposes, cf. ULAM (1929), TARSKI (1930b), STONE (1936). It is also closely connected with Theorem 3, because Lindenbaum algebras determined by consistent (by consistent and complete) sets of sentences are Boolean ideals (Boolean prime ideals). RASIOWA and SIKORSKI (1950) made use of the following modification of the theorem of extension of ideals to prime ideals.

LEMMA RS1. *If (i) a_0 is an element of a Boolean algebra B different from its unit element, and (ii) $a_i \in B$ for any $i \in I$ entails $\bigcup\{a_i \in B: i \in I\} \in B$, then there exists in B a prime ideal J such that $a_0 \in J$ and $[\bigcup\{a_i: i \in I\}]_J = \bigcup\{[a_i]_J: i \in I\}$, where $[a]_J$ stands for the equivalence class determined by the element a and the congruence induced by the prime ideal J .*

Note that A. Tarski made an observation simplifying the proof of Lemma RS1, cf. FEFERMAN (1952). The second lemma can be formulated as follows:

LEMMA RS2. *If (i) $X \subseteq S_1$, (ii) X/\approx is a prime ideal in S_1/\approx , (iii) $[\bigcup\{[A(x/t)]_\approx: t \in T_\omega\}]_{X/\approx} = \bigcup\{[A(x/t)]_\approx: t \in T_\omega\}_{X/\approx}$ for any formula $A(x)$ with one free variable x , and (iv) $[a_0]_\approx \in X/\approx$, then*

$$a_0 \notin E(\mathfrak{M}(X, S_1, T_\omega)).$$

A similar method of proving Theorem 2 was developed by L. RIEGER (1951).

IV. ON THE TERM "LINDENBAUM ALGEBRA"

The Lindenbaum algebras are called by some authors (cf. RIEGER, 1951; and HENKIN and TARSKI, 1961) the *Lindenbaum-Tarski algebras*. The question which terminology to adopt gave rise to some discussion which was reported in RASIOWA and SIKORSKI (1968).

The quotation below comes from A. TARSKI (1935, p. 510).

"The interpretation of the algebra of logic in sentential algorithm can clearly be modified so as to avoid the replacement of logical identity by another equivalence relation. To obtain this modification we consider, instead of sentences $x \in S$, the equivalence classes $X \subseteq S$ each of which consists of all sentences y which are equivalent, in the sense of Def. 4b, to some sentence x ($y \equiv x$). For these equivalence classes we define in an appropriate way the relations and operations \supset , $+$, ..., etc.; for instance, the formula ' $X \supset Y$ ' will express the fact that $x \supset y$ (in the sense of Def. 4a), for all $x \in X$ and $y \in Y$. The elements 0 and 1 are now interpreted 'effectively'—in fact 0 as the set of all sentences $x \in S$ such that $\bar{x} \in L$, and 1 simply as the set L ."

The following passage can be found in RASIOWA and SIKORSKI (1968, pp. 245–246):

"A. Lindenbaum was a prominent Polish mathematician who died prematurely (was killed by Nazis during the Second World War) and

whose various results were not published. To honour Lindenbaum, Polish logicians usually call $\mathfrak{A}(\mathfrak{T})$ the *Lindenbaum algebra*. This name for $\mathfrak{A}(\mathfrak{T})$ was also used by the authors of this book. We were also influenced by the fact that MCKINSEY (1941), a close collaborator of Tarski, called the method of treating the set of all formulas of a formalized language of the zero order 'an unpublished method of Lindenbaum' (p. 222, line 12) 'explained to me by Professor Tarski' (footnote 7). We have the duty to quote here also the following remark of HENKIN and TARSKI (1961, p. 85, footnote 4): 'Apart from a different treatment on the notion of equality, the quotient algebras F/\approx were explicitly introduced in TARSKI (1935, p. 510)', 'the historical justification given in RASIOWA and SIKORSKI (1958, p. 143, footnote 1), for calling these algebras Lindenbaum algebras seems to be incorrect'."

The concepts and symbols used in this section are explained in TARSKI (1935) and RASIOWA and SIKORSKI (1968).

References

- BETH, E. W., 1951, *A topological proof of the theorem of Löwenheim-Skolem-Gödel-Tarski*, Proceedings of the Royal Academy of Sciences, series A, vol. 52, pp. 536–444
 Epitah, 1945a, Annales de la Société Polonaise de Mathématique 18, p. 9
 Epitah, 1945b, Fundamenta Mathematicae, vol. 33, p. 5
- FRAENKEL, A. A., 1967, *Lebenskreise. Aus den Erinnerungen eines jüdischen Mathematikers* (Stuttgart)
- FRANKEL, A. A., and Y. Bar-Hillel, 1958, *Foundations of set theory* (North-Holland Publ. Co., Amsterdam)
- FEFERMAN, S., 1952, *Review of the paper: H. Rasiowa and R. Sikorski, "A proof of the completeness theorem of Gödel*, Fundamenta Mathematicae 37, (1950), pp. 193–200, Journal of Symbolic Logic, vol. 19, p. 72
- GÖDEL, K. 1930, *Die Vollständigkeit der Axiome der logischen Funktionenkalkuls*, Monatshefte für Mathematik und Physik, vol. 37, pp. 349–360
- GROMSKA, D., 1948, *Philosophes Polonais morts entre 1939 et 1943*, Studia Philosophica, vol. 3, pp. 31–97
- GRZEGORCZYK, A., 1969, *Zarys logiki matematycznej* (Outline of mathematical logic, Polish), Wydanie drugie (Warszawa)
- HASENJÄGER, G., 1953, *Bemerkung zu Henkin's Beweis für die Vollständigkeit des Prädikatenkalkuls der ersten Stufe*, Journal of Symbolic Logic, vol. 18, pp. 42–48
- HENKIN, L., 1949, *A proof of completeness for the first order junctional calculus*, Journal of Symbolic Logic, vol. 14, pp. 159–166
- HENKIN, L., 1950, *An algebraic characterization of quantifiers*, Fundamenta Mathematicae, vol. 37, pp. 63–74
- HENKIN, L., and A. TARSKI, 1961, *Cylindric algebras*, Proceedings of Symposia in Pure Mathematics. II, Lattice theory, pp. 83–113

- Łoś, J., 1949, *O matrycach logicznych* (On logical matrices, Polish), *Travaux de la Société des Sciences et des Lettres de Wrocław, Sér. B, Nr. 19* (Wrocław) 44 pp.
- Łoś, J., 1955, *The algebraic treatment of the methodology of elementary deductive systems*, *Studia Logica*, vol. 2, pp. 151–212
- ŁUKASIEWICZ, J., and A. TARSKI, 1930, *Untersuchungen über den Aussagenkalkül*, *Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie, Classe III*, vol. 23, pp. 30–50
- MAUTNER, F. I., 1946, *Logic as invariant theory, an extension of Klein's Erlangen Program*, *American Journal of Mathematics*, vol. 68, pp. 345–386
- MCKINSEY, J. C. C., 1941, *Solution of the decision problem for Lewis S2 and S4 systems with an application to topology*, *Journal of Symbolic Logic*, vol. 6, pp. 117–134
- MESCHKOWSKI, H., 1964, *Mathematiker-Lexikon* (Mannheim, Zürich)
- MOSTOWSKI, A., 1948, *Proofs of non-deductibility in intuitionistic functional calculus*, *Journal of Symbolic Logic*, vol. 13, pp. 204–207
- RASIOWA, H., and R. SIKORSKI, 1950, *A proof of the completeness theorem of Gödel*, *Fundamenta Mathematicae*, vol. 37, pp. 193–200
- RASIOWA, H., and R. SIKORSKI, 1968, *The mathematics of metamathematics* (PWN, Warszawa)
- RASIOWA, H., and R. SIKORSKI, 1958, *On isomorphism of Lindenbaum algebras with fields of sets*, *Colloquium Mathematicum*, vol. 5, pp. 143–158
- REICHBACH, J., 1955, *O pełności węższego rachunku funkcyjnego*, *Studia Logica*, vol. 2, pp. 213–228
- RIEGER, L., 1951, *On free \aleph_0 -complete Boolean algebras*, *Fundamenta Mathematicae*, vol. 38, pp. 35–52
- STONE, H. M., 1936, *The theory of representation for Boolean algebras*, *Transactions of the American Mathematical Society*, vol. 40, pp. 37–111
- SURMA, S. J., 1968, *Cztery studia z metamatematyki*, *Studia Logica*, vol. 23, pp. 80–114
- SURMA, S. J., 1969, *Some results in metamathematical non-effectiveness*, *Universitas Jagellonica Acta Scientiarum Litterarumque CCVIII, Schedae Logicae*, vol. 4, pp. 31–37
- SURMA, S. J., 1973, *The concept of the Lindenbaum algebra: its genesis*, in: *Studies in the History of Mathematical Logic*, ed. S. J. Surma, pp. 239–253 (Ossolineum, Wrocław)
- SZUMAKOWICZ, E., 1977, *Adolf Lindenbaum's contributions to contemporary logic and the foundations of mathematics*, Master's Thesis written under the supervision of S. J. Surma at the Department of Logic, Jagiellonian University of Cracow
- TARSKI, A., 1930a, *Über einige fundamentale Begriffe der Metamathematik*, *Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie, Classe III*, vol. 23, pp. 22–29
- TARSKI, A., 1930b, *Une contribution à la théorie de la mesure*, *Fundamenta Mathematicae*, vol. 15, pp. 42–50
- TARSKI, A., 1935, *Grundzüge des Systemenkalküls*, Erster Teil, *Fundamenta Mathematicae*, vol. 25, pp. 503–526
- ULAM, K., 1929, *Concerning functions of sets*, *Fundamenta Mathematicae*, vol. 14, pp. 231–233
- WAJSBERG, M., 1937, *Metalogische Beiträge*, *Wiadomości Matematyczne*, vol. 43, pp. 1–38

FREGE'S NOTION OF "BEDEUTUNG"

IGNACIO ANGELELLI

Austin, Texas, U.S.A.

The aim of this paper * is to contribute to a better understanding of Frege's notion of *Bedeutung*. The paper is divided into four parts. In 1 I quote the crucial texts, in 2, I present an analysis of them. In the analysis several problems arise, listed at the end of Section 2. In 3 I propose an interpretation which I apply to solve those problems. Some critical remarks are added in 4 by way of conclusion.

Abbreviations: BP₁ = the first principle of *Bedeutung*, BP₂ = the second principle of *Bedeutung*, V = the conjecture (*Vermutung*) V, RS = the rule of substitutivity of identicals. The letters *p*, *q*, *r* are used to abbreviate three statements concerning the relationship among *Bedeutung* of sentences, truth-value of sentences, and singular terms occurring in sentences. For all these see Section 2. *Principle B* or the *principle of Bedeutung* is explained in Section 3.

1. The texts¹

A1. *Bisher sind Sinn und Bedeutung nur von solchen Ausdrücken, Wörtern, Zeichen betrachtet worden, welche wir Eigennamen genannt haben. Wir fragen nun nach Sinn und Bedeutung eines ganzen Behauptungssatzes.*

* I am grateful to the Committee on Attendance at Meetings of Learned Societies and to The Arts and Sciences Foundation of The University of Texas at Austin, for helping me to attend the 6th International Congress of Logic, Methodology and Philosophy of Science.

¹ All the quoted texts in this section are from FREGE SUB, pp. 32–36.

- A2. Ein solcher Satz enthält einen Gedanken (Ich verstehe unter Gedanken [...] dessen objektiven Inhalt [...]). Ist dieser Gedanke nun als dessen Sinn oder als dessen Bedeutung anzusehen?
- A3. Nehmen wir einmal an, der Satz habe eine Bedeutung!
- A4. Ersetzen wir nun in ihm ein Wort durch ein anderes von derselben Bedeutung, aber anderm Sinne, so kann dies auf die Bedeutung des Satzes keinen Einfluss haben.
- A5. Nun sehen wir aber, dass der Gedanke sich in solchem Falle ändert; denn es ist z.B. der Gedanke des Satzes "der Morgenstern ist ein von der Sonne beleuchteter Körper" verschieden von dem des Satzes "der Abendstern ist ein von der Sonne beleuchteter Körper". Jemand, der nicht wüsste, dass der Abendstern der Morgenstern ist, könnte den einen Gedanken für wahr, den andern für falsch halten.
- A6. Der Gedanke kann also nicht die Bedeutung des Satzes sein,
- A7. vielmehr werden wir ihn als den Sinn aufzufassen haben.
- B1. Wie ist es nun aber mit der Bedeutung? Dürfen wir überhaupt danach fragen? Hat vielleicht ein Satz als Ganzes nur einen Sinn, aber keine Bedeutung? Man wird jedenfalls erwarten können, dass solche Sätze vorkommen, ebensogut, wie es Satzteile gibt, die wohl einen Sinn, aber keine Bedeutung haben.
- B2. Und Sätze, welche Eigennamen ohne Bedeutung enthalten, werden von der Art sein. Der Satz "Odysseus wurde tief schlafend in Ithaka ans Land gesetzt" hat offenbar einen Sinn. Da es aber zweifelhaft ist, ob der darin vorkommende Name "Odysseus" eine Bedeutung habe, so ist es damit auch zweifelhaft, ob der ganze Satz eine habe.
- B3. Aber sicher ist doch, dass jemand, der im Ernst den Satz für wahr oder für falsch hält, auch dem Namen "Odysseus" eine Bedeutung zuerkennt, nicht nur einen Sinn;
- B4. denn der Bedeutung dieses Namens wird ja das Prädikat zu- oder abgesprochen. Wer eine Bedeutung nicht anerkennt, der kann ihr ein Prädikat weder zu- noch absprechen.
- B5. Nun wäre aber das Vordringen bis zur Bedeutung des Namens überflüssig; man könnte sich mit dem Sinne begnügen, wenn man beim Gedanke stehenbleiben wollte. Käme es nur auf den Sinn des Satzes,

den Gedanken, an, so wäre es unnötig, sich um die Bedeutung eines Satzteils zu kümmern;

- B6. *für den Sinn des Satzes kann ja nur der Sinn, nicht die Bedeutung dieses Teils in Betracht kommen. Der Gedanke bleibt derselbe, ob der Name "Odysseus" eine Bedeutung hat oder nicht.*
- B7. *Dass wir uns überhaupt um die Bedeutung eines Satzteils bemühen, ist ein Zeichen dafür, dass wir auch für den Satz selbst eine Bedeutung im allgemeinen anerkennen und fordern.*
- B8. *Der Gedanke verliert für uns an Wert, sobald wir erkennen, dass zu einem seiner Teile die Bedeutung fehlt.*
- B9. *Wir sind also wohl berechtigt, uns nicht mit dem Sinne eines Satzes zu begnügen, sondern auch nach seiner Bedeutung zu fragen.*
- B10. *Warum wollen wir denn aber, dass jeder Eigenname nicht nur einen Sinn, sondern auch eine Bedeutung habe? Warum genügt uns der Gedanke nicht? Weil und soweit es uns auf seinen Wahrheitswert ankommt.*
- B11. *Nicht immer ist dies der Fall. Beim Anhören eines Epos z.B. fesseln uns neben dem Wohlklange der Sprache allein der Sinn der Sätze und die davon erweckten Vorstellungen und Gefühle. Mit der Frage nach der Wahrheit würden wir den Kunstgenuss verlassen und uns einer wissenschaftlichen Betrachtung zuwenden.*
- B12. *Daher ist es uns auch gleichgültig, ob der Name "Odysseus" z.B. eine Bedeutung habe, solange wir das Gedicht als Kunstwerk aufnehmen.*
- B13. *Das Streben nach Wahrheit also ist es, was uns überall vom Sinne zur Bedeutung vorzudringen treibt.*
- B14. *Wir haben gesehen, dass zu einem Satze immer dann eine Bedeutung zu suchen ist, wenn es auf die Bedeutung der Bestandteile ankommt; und das ist immer dann und nur dann der Fall, wenn wir nach dem Wahrheitswerte fragen.*
- B15. *So werden wir dahin gedrängt, den Wahrheitswert eines Satzes als seine Bedeutung anzuerkennen.*
- C1. *Wenn unsere Vermutung richtig ist, dass die Bedeutung eines Satzes sein Wahrheitswert ist, so muss dieser unverändert bleiben, wenn ein Satzteil durch einen Ausdruck von derselben Bedeutung, aber anderm Sinne ersetzt wird.*

- C2. *Und das ist in der Tat der Fall. Leibniz erklärt gradezu: “Eadem sunt, quae sibi mutuo substitui possunt, salva veritate”.*
- C3. *Was sonst als der Wahrheitswert könnte auch gefunden werden, das ganz allgemein zu jedem Satze gehört, bei dem überhaupt die Bedeutung der Bestandteile in Betracht kommt, was bei einer Ersetzung der angegebener Art unverändert bliebe?*
- D1. *Es soll nun die Vermutung, dass der Wahrheitswert eines Satzes dessen Bedeutung ist, weiter geprüft werden.*
- D2. *Wir haben gefunden, dass der Wahrheitswert eines Satzes unberührt bleibt, wenn wir darin einen Ausdruck durch einen gleichbedeutenden ersetzen: wir haben aber dabei den Fall noch betrachtet, dass der zu ersetzende Ausdruck selber ein Satz ist.*
- D3. *Wenn nun unsere Ansicht richtig ist, so muss der Wahrheitswert eines Satzes, der einen andern als Teil enthält, unverändert bleiben, wenn wir für den Teilsatz einen andern einsetzen, dessen Wahrheitswert derselbe ist.*

2. Analysis of the texts

In the initial pages of SUB, that is, up to text A1, Frege explains the notion of *Bedeutung* only for the special case of singular terms: *Die Bedeutung eines Eigennamens ist der Gegenstand selbst, den wir damit bezeichnen* (p. 30).²

As soon as we reach text A1 we observe that Frege has more ambitious plans: he wants to extend the notion of *Bedeutung* from singular terms to sentences. In A2 Frege refers to the *Gedanke* (thought) as the *Inhalt* (content) of the sentence. The question arises whether the *Gedanke* is the *Bedeutung* of the sentence. Any unprejudiced reader, who imagines that Frege is looking for an entity that stands to the sentence in the same relation in which objects stand to their singular terms, will probably guess

² Cf. similar texts: *wir nennen den Gegenstand, den ein Eigenname bezeichnet, seine Bedeutung*, Nachlass I, p. 208; *der Mond selbst ist die Bedeutung des Ausdrückes ‘der Mond’*, Nachlass I, p. 275 (cf. *ich vergleiche den Mond selbst mit der Bedeutung*, SUB, p. 30, *der Mond selbst (d. h. die Bedeutung des Wortes ‘Mond’)*, Nachlass II, p. 245); *kann nicht der Berg Aetna selbst sein, kann nicht die Bedeutung dieses Namens sein* (Nachlass II, p. 127); *die Bedeutung des Namens als dasjenige, von dem etwas ausgesagt wird*, (Nachlass II, p. 128); *der Gegenstand, von dem ich etwas aussage [...] ist immer die Bedeutung des Zeichens* (Nachlass II, p. 231).

that the *Gedanke* is the *Bedeutung* of the sentence. In A6, however, Frege affirms that this is not possible. He presents this assertion as a conclusion from three premisses. The first (A3) is the assumption that the sentence has a *Bedeutung*. The second premiss (A4) says that substitution of words of equal *Bedeutung* inside a sentence cannot affect the *Bedeutung* of the sentence. The third premiss (A5) claims that the substitution of terms of equal *Bedeutung* in a sentence does not preserve the identity of the *Gedanke* expressed by the sentence.

We may ignore the third premiss (A5) as well as A7, which have to do with *Sinn*. The second premiss (A4) is the one that is very important for us. Just as it stands, it is a statement about the *Bedeutung* of sentences. Frege, however, is not supposed to know anything at this point about the *Bedeutung* of sentences, except what follows from the notion of *Bedeutung* in general. Thus we have to construe the second premiss as an application to the particular case of sentences of some general principle for *Bedeutung* that Frege has in mind but that he has not written down. Let us call this the *first principle of Bedeutung* (BP₁):

*if two expressions E and E' have the same Bedeutung, and E occurs in an expression A(E) which has a Bedeutung, then the result of substituting E' for E, A(E'), has the same Bedeutung as A(E).*³

Why should we accept BP₁, what is its justification or indeed its meaning? Clearly, for a reader who only knows about *Bedeutung* Frege's explanations of the *Bedeutung* of singular terms in the initial pages of SUB, BP₁ makes hardly any sense at all. This is our *problem 1*.

Next we have the series of texts B, whose main objective is to establish the conjecture (*Vermutung*, cf. texts C and D) that the *Bedeutung* of a sentence is identical to its truth-value. Frege's presentation of his reasoning is not up to his reputation; some benevolent reconstruction

³ BP₁ reappears in other writings: *Die Bedeutung eines Satzes muss etwas sein, was bestehen bleibt, wenn einer seiner Teile durch etwas Gleichbedeutendes ersetzt wird* (Nachlass I, p. 251); *wenn man in einem Satze oder Satzteile einen Bestandteil durch einen gleichbedeutenden [...] hat der abgeänderte Satz oder Satzteil dieselbe Bedeutung wie der ursprüngliche* (Nachlass I, p. 276); *wenn der Gedanke die Bedeutung des Satzes wäre, so änderte er sich nicht, wenn einer seiner Teile ersetzt würde durch einen anderen Ausdruck von derselben Bedeutung* (Nachlass II, p. 235); *die Bedeutung des Satzes muss etwas sein, was sich nicht ändert, wenn wir ein Zeichen durch ein anderes ersetzen, das dieselbe Bedeutung [...] hat (ibid.); da nun die Sätze in den Bedeutungen ihrer Bestandtheile vollkommen übereinstimmen, müssen auch sie dieselbe Bedeutung haben* (Nachlass II, p. 245).

is needed. I propose the following. Let us consider these three statements about an arbitrary sentence s :

- p) every singular term in s has a *Bedeutung*,
- q) the sentence s has a truth-value,
- r) the sentence s has a *Bedeutung*.

In B1 Frege says that r does not necessarily hold of all sentences: there may be sentences without *Bedeutung*—just as there are singular terms without it. In B2 we find (1) $\neg p \rightarrow \neg r$. In B3 we are told that (2) $q \rightarrow p$, which is supported by the remark in B4 that there can be no prediction if no object is there. B5 may be analyzed as (3) $\neg r \rightarrow \neg p$. Leaving aside B6, which has to do with *Sinn*, we find in B7: (4) $p \rightarrow r$. B8 is vague—what does ‘Wert’ mean? *Wahrheitswert* or *Erkenntniswert*?⁴ In B9 Frege feels that he can draw the conclusion that it is justified to look for the *Bedeutung* of sentences. In B10 we have two questions: (i) why do we want *Bedeutung* of singular terms? (ii) why do we want *Bedeutung* of sentences? Both questions are answered at once: we want *Bedeutung* exactly to the extent that we are concerned with the truth-value of the *Gedanke*. This might be expanded into the following: (5) $p \leftrightarrow q$, (6) $r \leftrightarrow q$. In B11 it is pointed out that we are not always interested in the truth-value of sentences. When we are not, B12 tells us that we are not interested in the *Bedeutung* of the singular terms either: (7) $\neg q \rightarrow \neg p$. All this indicates that there is a close relationship between truth and *Bedeutung* (B13). B14 summarizes in two parts the preceding considerations: (8) $p \rightarrow r$, (9) $p \leftrightarrow q$. Apparently, the conditional $r \rightarrow p$ is left out.⁵ Finally, B15 formulates the thesis that the truth-value of a sentence is the same as its *Bedeutung*. We are forced (*gedrängt*) to accept this thesis, if not as a demonstrated truth, in any event as a conjecture, *Vermutung*, cf. texts C and D. The only reason why Frege continues his paper beyond

⁴ The two words occur, for example, in the last two paragraphs of SUB.

⁵ Curiously, in *Nachlass* I, pp. 250–251 Frege repeats the same reasoning of our texts B1–B15 and again seems to overlook the $r \rightarrow p$ half of $p \leftrightarrow r$, although it is obviously needed: *Wenn es uns also darauf ankommt, dass der Eigenname ‘Aetna’ etwas bezeichne, wird es uns auch auf die Bedeutung des ganzen Satzes ankommen. Dass der Name ‘Aetna’ etwas bezeichne, ist uns aber immer dann und nun dann von Wert, wenn es uns auf die Wahrheit im wissenschaftlichen Sinne ankommt. Eine Bedeutung wird also unser Satz dann und nur dann haben, wenn der in ihm ausgedrückte Gedanke wahr oder falsch ist.* These three sentences are $p \rightarrow r$, $p \leftrightarrow q$ and $r \leftrightarrow q$, respectively. The last is presented as a conclusion from the first two, which obviously requires the additional $r \rightarrow p$.

B15 is to increase the plausibility of this conjecture, to which I will refer as conjecture V.

In analyzing the sequence B1–B15 we cannot fail to recognize in the biconditional $p \leftrightarrow r$ an application, to the particular case of sentences, of another general principle of *Bedeutung* which Frege, again, takes for granted rather than presenting properly. This second principle of *Bedeutung* may be called the principle of 'existence' of *Bedeutung*: *all the expressions occurring as parts of a complex expression have a Bedeutung iff the complex expression has a Bedeutung*.⁶ No reader of SUB can be expected to make any sense of this principle on the basis of Frege's explanations concerning the *Bedeutung* of singular terms at the beginning of the paper (*problem 2*).

There are further difficulties in the sequence B1–B15. First, the conjecture V itself is hard to understand. For readers who, again, about *Bedeutung* only know that the *Bedeutung* of a singular term is the object denoted by it, it is surely very awkward to learn that the *Bedeutung* of a sentence is its truth-value (*problem 3*). To the reader's distress, Frege insists elsewhere that the truth-value is the *Bedeutung* of the sentence just like (*ebenso wie*) for example number 4 is the *Bedeutung* of '2+2' or London is the *Bedeutung* of 'the capital of England' (FUB, pp. 13, 16⁷).

Additional difficulties in the sequence B1–B15 are also related to the conjecture V. How is V suggested by texts B1–B14? The closest to V in these texts is the biconditional $r \leftrightarrow q$. Thus it seems appropriate to view the entire sequence B1–B14 as oriented towards the establishment of $r \leftrightarrow q$ as a conclusion from the two premisses $p \leftrightarrow r$ and $p \leftrightarrow q$.⁸ But why should $r \leftrightarrow q$ suggest, let alone imply V? (*problem 4*). Moreover, why a mere conjecture and why not a full assertion or stipulation? If V has to be proved, how can it be proved? (*problem 5*).

Let us now approach texts C and D. Each of them represents an attempt to confirm the conjecture V by showing that a sentence implied by V is true.

⁶ Cf. *Nachlass I*, p. 262: *Jeder dieser Teile muss ebenfalls eine Bedeutung haben, wenn der ganze Satz eine Bedeutung [...] haben soll.* *Nachlass I*, p. 211: *Wenn man den Satz in Teile zerlegen kann, von denen jeder bedeutungsvoll ist, so hat auch der Satz eine Bedeutung.*

⁷ The sentence $3^2 = 4$ bedeutet the truth-value false gerade so, wie '2^a' die Zahl Vier bedeutet, GRG I § 2. Russell could not make sense of this theory, cf. his letter in *Nachlass II*, p. 233.

⁸ This trend of reasoning may be observed in other passages, such as, for example, *Nachlass I*, pp. 250–251 (see footnote 5). Related texts are in *Nachlass II*, pp. 235, 240, 247, *Nachlass I*, pp. 210–211.

In C1 Frege asserts a conditional whose antecedent is the conjecture V and whose consequent is the thesis that singular terms of equal *Bedeutung* are interchangeable *salva veritate*. In C2 Frege claims that this thesis is true. The thesis is often formulated as a rule: the rule of 'substitutivity of identicals' (cf. ANGELELLI, 1976). Let me use the abbreviation RS to refer to the thesis or to the rule, as the context may require. Thus the conditional asserted by Frege in C1 is: $V \rightarrow RS$. Because of believing that RS is true (C2), for Frege the conditional $V \rightarrow RS$ is true, but this makes it only 'materially' true. By adding the principle BP_1 , we obtain the logically true conditional: $BP_1 \wedge V \rightarrow RS$. BP_1 says that the substitution of singular terms of equal *Bedeutung* does not touch the *Bedeutung* of the sentence in which they occur; RS says the same, except that it has 'truth-value of the sentence' instead of '*Bedeutung* of the sentence'; V secures the identity of truth-value and *Bedeutung* of sentences.

In C3 Frege states his belief that there is no other entity available to perform the duties of *Bedeutung* of a sentence apart from the truth-value of the sentence.

In D1 Frege announces his plan of testing V a second way. D2 restates RS while pointing out that we have not yet considered the case where sentences are substituted for one another. In D3 we have a conditional again: if the *Bedeutung* of a sentence is identical to its truth-value, then substitution of a sentence by another of same truth-value inside a compound sentence should not change the truth-value of the compound.

As it stands, this conditional is not logically true. Addition of BP_1 to the antecedent gives a logically true conditional: the consequent becomes an instance of BP_1 if 'truth-value' is replaced by '*Bedeutung*', i.e. if the conjecture V is used. The remaining pages of Frege's paper aim at proving if not the truth at least the plausibility of the consequent (*mit hinreichender Wahrscheinlichkeit*, p. 49).

To sum up, let me list the difficulties encountered in the analysis of the texts: (1) How can we make sense of BP_1 ? (2) How can we make sense of BP_2 ? (3) How can we understand the Fregean doctrine that truth-values stand to sentences like objects to singular terms? (4) How does $r \leftrightarrow q$ help to conjecture the identity of truth-value and *Bedeutung*? (5) Why does Frege restrict himself to a mere conjecture and how is the conjecture to be proved?

3. Interpretation

As a first step in the interpretation I will distinguish two meanings of the term 'Bedeutung': (i) *Bedeutung* in the semantical sense, semantic *Bedeutung* (for example, the *Bedeutung* of the singular term 'Caesar' is the object Caesar), (ii) *Bedeutung* in the sense of importance, *Bedeutung-importance* (for example, the *Bedeutung* of Caesar is enormous). Of course, in the case of *Bedeutung-importance* one has to explain *relative to what* is the importance understood⁹.

The distinction is not made by Frege when he uses the word technically (from the time of SUB on) but the two meanings are recognizable in his pre-technical use of the term.¹⁰ To show this, as well as to suggest how the Fregean notion of *Bedeutung-importance* is to be approached, I will quote a few texts from the *Begriffsschrift*:

- (1) *Deshalb ist auf den Ausdruck alles dessen verzichtet worden, was für die Schlussfolge ohne Bedeutung ist. Ich habe das, worauf allein es mir ankam, in § 3 als begrifflichen Inhalt bezeichnet* (BG, Vorwort).¹¹

I have distinguished the two meanings of 'Bedeutung' in my dissertation *Studies on G. Frege and traditional philosophy* (Univ. of Fribourg, 1965), Chapter 2, published with the same title by Reidel in 1967, cf. especially Section 2.26, in which I formulated essentially the principle that in this paper I call principle B of *Bedeutung*. Subsequently a similar interpretation of *Bedeutung* as importance was presented by TUGENDHAT (1970). Although I am of course sympathetic with his approach, there is at least one serious discrepancy that I should mention here. Tugendhat's definition of *Wahrheitspotential* (*Bedeutung-importance*) contains a clause, superfluous in my view, that in the special case of sentences makes the truth-value of a sentence identical *per definitionem* with the *Bedeutung* or *Wahrheitspotential* of the sentence. I find this as interpretation of Frege inappropriate, because it transforms Frege's conjecture V into a definition! Tugendhat's definition of *Wahrheitspotential* is the following: "two expressions φ and ψ have the same truth-value potential if and only if, whenever each is completed by the same expression to form a sentence, the two sentences have the same truth-value" (p. 180). For the special case in which φ and ψ are sentences, say p and q , Tugendhat argues that "they are not susceptible to being completed as sentences by a further expression. Therefore the addition 'whenever each is completed...' is superfluous in this case, and the definition is reduced to the simple form: two sentences p and q have the same truth-value potential if and only if they have the same truth-value". (p. 180).

¹⁰ Frege refers to SUB as the first paper where the term 'Bedeutung' occurs in a technical sense (*Nachlass II*, p. 41, p. 96).

¹¹ For a related text cf. *Nachlass I*, p. 37, footnote: *eine Verschiedenheit nur dann einen logischen Wert hat, wenn sie die möglichen Schlussfolgerungen berührt.*

- (2) *Die Inhalte von zwei Urtheilen in doppelter Weise verschieden sein können: erstens so, dass die Folgerungen, die aus dem einen in Verbindung mit bestimmten andern gezogen werden können, immer auch aus dem zweiten in Verbindung mit denselben andern Urtheilen folgen; zweitens so, dass dies nicht der Fall ist. Die beiden Sätze: "bei Plataeae siegten die Griechen über die Perser" und "bei Plataeae wurden die Perser von den Griechen besiegt" unterscheiden sich in der ersten Weise. Wenn man nun auch eine geringe Verschiedenheit des Sinnes erkennen kann, so ist doch die Uebereinstimmung überwiegend. Ich nenne nun denjenigen Theil des Inhaltes, der in beiden derselbe ist, den begrifflichen Inhalt. Da nur dieser für die Begriffsschrift von Bedeutung ist, so braucht sie keinen Unterschied zwischen Sätzen zu machen, die denselben begrifflichen Inhalt haben (BG, § 3).*
- (3) *So ist denn mit der Einführung eines Zeichens der Inhaltsgleichheit nothwendig die Zwiespältigkeit in der Bedeutung aller Zeichen gegeben, indem dieselben bald für ihren Inhalt, bald für sich selber stehen" (BG, § 8, cf. GRG II, § 98).*
- (4) *Es bedeute nun [...] $A \equiv B$: das Zeichen A und das Zeichen B haben denselben begrifflichen Inhalt, sodass man überall an die Stelle von A B setzen kann und umgekehrt (BG, § 8)¹².*

The meaning of ‘Bedeutung’ in (1) and (2) is clearly different from that in (3). In the latter, *Bedeutung* is used in the semantical sense (*für etwas stehen*, to stand for something), in (1) and (2) *Bedeutung* is used in the sense of importance—specifically, importance relative to logical inference.

From (2) we learn that the conceptual conteut, i.e. what is important or ‘has *Bedeutung*’ in the content of an expression E —we may say: the *Bedeutung*-importance of ‘ E ’—is to be reached by *abstraction*. We have to abstract from anything in the content of E that is not included in the content of any other expression E' that stands to E in a certain relation \sim . This relation \sim is in the texts (1) and (2) something like ‘behaving equally in inferences’. If in the expressions E , E' (for example ‘aber’ and ‘und’) we abstract from anything that is not invariant with respect to \sim , the conceptual content of E and the conceptual content of E' become identical.

For the analysis of Frege’s notion of *Bedeutung* in the period in which

¹² Note that in this rule at the end of § 8 the sign \equiv means identity of *conceptual* content whereas in the preceding discussion within § 8 Frege examines the possibility of introducing a symbol to express identity of content *simpliciter* (*Inhaltsgleichheit*). A full analysis of this curious switch cannot be carried out here.

he used the word as a technical term (from SUB on) it seems convenient to make the abstraction that leads to *Bedeutung*-importance relative to the relation of interchangeability *salva veritate* rather than to 'behaving equally in inferences'. In other words, it seems convenient to conceive the *Bedeutung*-importance specifically as importance for truth rather than as importance for inferences. This choice not only seems to work but it is justified by text (4), at least in the sense that in text (4) Frege takes equal importance for logical inferences to entail equal importance for truth (substitutivity *salva veritate*).¹³

Let me then propose, as starting point of my interpretation, the following statement: *The Bedeutung-importance of an expression E = the Bedeutung-importance of an expression E' iff E and E' are interchangeable salva veritate.* I will call this the *principle of Bedeutung*, briefly *principle B*.

Although principle B is suggested by the *Begriffsschrift*,¹⁴ it does not seem to occur in Frege's writings. Still, it is a very Fregean principle. To show this, let us represent its general form by $\tilde{a} = \tilde{b}$ iff $a \sim b$, where \sim is a relation defined on the domain of objects a, b, \dots . If the objects are predicates and \sim is the relation of being true of the same individuals, then a is the *Wertverlauf* of a and the principle becomes Axiom V of *Grundgesetze*. If the objects a, b, \dots , are again predicates but now \sim is the relation of being *gleichzahlig* (equinumerous, GRL, § 68), then \tilde{a} is the number of a and the principle becomes another fundamental thesis of Frege's philosophy (GRL, § 62 ff).

In the given formulation of principle B I have not made the interchangeability *salva veritate* relative¹⁵ to any specific set of sentences;

¹³ Frege describes what is relevant to logic not only as 'what matters for inference' but also as 'what has to do with truth': *Die Logik betrachtet ihre Gegenstände sofern sie wahr sind*, *Nachlass I*, pp. 2, 3. In GED: *so weist 'wahr' [...] der Logik die Richtung* (p. 58). One might speculate that the two characterizations of logic in terms of inference and in terms of truth amount to the same, given that good inferences are defined as (formally) truth-preserving. Frege himself, for example, views the active-passive transformations *both* as not affecting inferences (BG, text (2)) *and* as not touching 'what is true or false' (GED p. 64).

¹⁴ Principle B is certainly akin to Frege's principle that words have *Bedeutung* only in context (GRL). In ANGEELL. (1967), 2.7, last paragraph I have referred to the relationship between this principle and the so-called "definitions by abstraction".

¹⁵ This relativity is exemplified by Leibniz, when he observes, in connection with the terms 'triangulum' and 'trilaterum' that *in omnibus propositionibus ab Euclide demonstratis de Triangulo substitui potest Trilaterum et contra salva veritate* (p. 236).

Frege, in fact, envisages the totality of language¹⁶; his notion of *Bedeutung* is, so to speak, 'absolute'.¹⁷

Principle BP₁, read with 'Bedeutung-importance' becomes trivially true under the assumption of principle B. Referring to the formulation of principle BP₁ in Section 2, suppose that the expressions *E* and *E'* have the same *Bedeutung* and assume principle B. Consider any statement *C(A(E))* in which *A(E)* occurs: it will have the same truth-value as *C(A(E'))* by principle B. But this amounts to saying that *A(E)* and *A(E')* have the same *Bedeutung* as well, because of *C* being any (*solution of problem 1*).

In principle BP₂ the phrase 'having a *Bedeutung*' means the same as 'standing in the relation of interchangeability *salva veritate* with some expression'. In other words, an expression *E* has a *Bedeutung* iff there is an expression *X* such that *X* and *E* are interchangeable *salva veritate*. Accordingly, principle BP₂ becomes the following: *all the expressions that are parts of a compound expression belong to the field of the relation of interchangeability *salva veritate* iff the compound expression belongs to that field.*

Atomic expressions do not belong to the field of the interchangeability *salva veritate* relation just by virtue of their 'shape'. To be admitted, they must fulfil certain conditions. Atomic singular terms must have a semantic *Bedeutung*, otherwise any sentence in which they occur lacks a truth-value (cf. text B4) and no substitution *salva veritate* is possible. 'Odysseus', for example, is not admissible, it is not interchangeable *salva veritate* with any expression (of course, not even with itself), it lacks *Bedeutung* (-importance), because it lacks semantic *Bedeutung*. Atomic predicates must have the property of being 'sharply defined' (*scharf begrenzt*, for every object *x*, either *x* falls under the predicate or not), otherwise any sentence in which they occur lacks 'Sinn' (GRG II, pp. 69–70), hence evidently cannot have a truth-value and again there is no possibility of substitution *salva veritate*. The predicates 'Haufe' ('heap', BG, p. 64) and

¹⁶ Cf. his remarks towards the end of SUB, where he says in connection with the difficulties involved in the second test of the conjecture V: *es ist schwer, alle in der Sprache gegebenen Möglichkeiten zu erschöpfen* (p. 49). An exceptional reference to interchangeability relative to specific sets of sentences ('mathematical compounds') in GEFÜGE, last page.

¹⁷ A possible hint at a relative *Bedeutung* in HUSS, pp. 319–320: *Für den Mathematiker*, who is interested in *die Sache selbst in die Bedeutung der Worte*, equiextensional concepts are 'equal'.

'Christ' ('christian', GRG II, p. 69) seem to be for Frege examples of non-admissible predicates. To be admissible (*zulässig*), i.e. to be *bedeutungsvoll*, a predicate must be sharply defined (GRG II, p. 77). I would add (without textual proof) that these necessary conditions on atomic expressions (singular terms and predicates) are also sufficient in order to have a *Bedeutung*-importance.

Frege, in connection with his system of *Grundgesetze*, assumes that the atomic expressions are *bedeutungsvoll*, have a *Bedeutung*-importance or, equivalently, in our interpretation, that they are interchangeable *salva veritate* with some expression. Having assumed this, he concentrates on the proof of the left-to-right half of BP_2 , namely that the rules of formation of his system preserve the property of having a *Bedeutung*-importance (GRG I, § 28–32). The other half of BP_2 , i.e. if a component lacks *Bedeutung*, then the compound lacks *Bedeutung* as well, follows from the existence of *Bedeutung* for atomic expressions in conjunction with the demonstrated left-to-right half of BP_2 : atomic components have *Bedeutung*; compound components, if well-formed, have *Bedeutung* too.

While these considerations represent a clarification of the meaning of BP_2 (*solution of problem 2*), the truth of BP_2 depends of course on the particular system relative to which it is claimed; for example, Frege's attempt to prove it for *Grundgesetze* has been regarded as unsuccessful (cf. THIEL, 1965).

Our clarification of the meaning of BP_2 requires taking '*Bedeutung*' in the importance sense. This seems to create a new problem concerning the biconditionals $p \leftrightarrow r$, $q \leftrightarrow p$ from texts B1–B15, used by Frege to infer $r \leftrightarrow q$. The biconditional $p \leftrightarrow q$, to be intelligible, needs '*Bedeutung*' in the semantic sense (it is because of 'Odysseus' standing for nothing that no sentence with 'Odysseus' is true or false) but $p \leftrightarrow r$, insofar as it is an instance of BP_2 must have '*Bedeutung*' in the importance sense, both in r and in p . Thus the argument $p \leftrightarrow r$, $q \leftrightarrow p$ therefore $r \leftrightarrow q$ seems to lose the uniqueness of its 'middle term' p . This difficulty is overcome by considering that Frege 'identifies' (in a sense to be explained below) the *Bedeutung*-importance and the semantic *Bedeutung* of singular terms.

Before moving to the solution of the remaining problems I would like to mention another feature of the notion of *Bedeutung*-importance as explicated by principle B. Two expressions cannot have the same *Bedeutung*-importance if their mutual substitution does not preserve wellformedness (this is an obvious prerequisite for the preservation of truth). It is not clear that anything analogous should necessarily apply to semantic *Be-*

deutung. Thus, the following text, obscure and unconvincing with semantic *Bedeutung*, becomes perfectly intelligible with *Bedeutung-importance*:

Die Worte ‘der Begriff Quadratwurzel aus Vier’ verhalten sich [...] in Hinsicht auf ihre Ersetzbarkeit wesentlich anders als die Worte ‘eine Quadratwurzel aus Vier’ [...] d.h. die Bedeutungen dieser beiden Wortverbindungen sind wesentlich verschieden (BGGE, p. 201).

The solution of the problems (3), (4) and (5) requires some further speculation on the nature of the principle B. We have reached the latter starting from the hints given in the *Begriffsschrift*. According to these, the *Bedeutung-importance* of an expression is an abstractum and the principle B is a theorem about abstract entities. Unfortunately, this is not the way in which Frege conceives and makes use of statements of the form of principle B, after the *Begriffsschrift* and within the program of what he calls *eine sehr ungewöhnliche Art der Definition* (GRL, § 63)¹⁸.

Let me again represent statements of the form of principle B by the schema $\tilde{a} = \tilde{b}$ iff $a \sim b$. Frege proceeds in two stages; I have referred to this procedure (sometimes wrongly called ‘definition by abstraction’) as the ‘looking around method’ (ANGELELLI, 1979). First, Frege stipulates (GRL, § 62–67) $\tilde{a} = \tilde{b}$ iff $a \sim b$ without assigning any denotation to the singular terms ‘ \tilde{a} ’, ‘ \tilde{b} ’, ... Secondly, he looks around (CARNAP’s phrase, 1956, p. 1) for entities that are *suitable* to become such a denotation—suitable in the sense of being compatible with the stipulation made in stage one. All suitable entities are equally eligible to become denotata of the symbols ‘ \tilde{a} ’, ‘ \tilde{b} ’.

Frege does not emphasize, as Carnap does, the freedom of choice of entities, limited only by the requirement of being ‘suitable’. He chooses the so-called ‘equivalence classes’ in the case of number (GRL, § 68) and perhaps in general (*Nachlass* II, p. 195–196). It is clear, however, from his considerations in GRL, § 69 that there is for him nothing necessary about the equivalence classes. As a matter of fact, all he wants from the equivalence classes is that they satisfy the general condition $\tilde{a} = \tilde{b}$ iff $a \sim b$; other properties of the equivalence classes are rather a hindrance (GRL, § 69, especially last paragraph). Moreover, he remarks, in con-

¹⁸ Surely GRG II, § 146 may be read again in the sense of abstraction. But I do not think Frege has been sensitive to this. He heard of ‘abstraction’ from PEANO (1896, *Nachlass* II, p. 192) but in his reply he insists on choosing the equivalence class, i.e. he continues to proceed according to the looking-around method (pp. 195–196).

nexion with the choice of the equivalence class as denotatum of the phrase 'the number of...', the following: *Ich lege [...] auf die Heranziehung des Umfangs eines Begriffes kein entscheidendes Gewicht* (GRL, § 107) and in the particular case of *Wertverlauf* he chooses truth-values rather than equivalence classes (GRG I, § 10).

Within the looking-around method, principle B is thus a stipulation to be made at the start, *before* having assigned any denotation to the phrases of the form 'the Bedeutung-importance of (an expression) E'. This denotation has to be secured next, by choosing from any of the entities shown to be suitable.

For Frege, the semantic *Bedeutung* of singular terms is suitable to become their *Bedeutung*-importance, if we only assume that he holds, in addition to RS (text C2, cf. also ANGELELLI, 1976) its converse. RS, in fact, amounts to saying that if two singular terms have the same semantic *Bedeutung*, they have the same *Bedeutung*-importance as well.

In the interpretation proposed here, Frege chooses the semantic *Bedeutung* of singular terms as their *Bedeutung*-importance.

Frege conjectures that the truth-value of a sentence is suitable to become the *Bedeutung*-importance of the sentence and he is inclined to regard the extension as suitable to become the *Bedeutung*-importance of predicates¹⁹, but this project is blocked by his peculiar theory of saturated vs. unsaturated symbols and entities²⁰, in such a way that it is not even possible to talk of 'the' *Bedeutung* of a predicate (*Nachlass I*, p. 275).

Turning now to the remaining problems, we find first that it is no longer awkward to hear from Frege that truth-values are the *Bedeutung* of sentences just like objects are the *Bedeutung* of singular terms. In fact, what Frege means is that the truth-values are suitable to perform the duties of *Bedeutung*-importance of sentences just like objects (the semantic

¹⁹ Frege affirms that *was zwei Begriffswörter bedeuten ist dann und nur dann dasselbe, wenn die zugehörigen Begriffsumfänge zusammenfallen* (*Nachlass I*, p. 133). This 'bedeuten' is certainly of the importance type, as confirmed by Frege's preceding remarks: *unbeschadet der Wahrheit, in jedem Satze Begriffswörter einander vertreten können, wenn ihnen derselbe Begriffsumfang entspricht, dass also auch in Beziehung auf das Schliessen und für die logische Gesetze Begriffe nur insofern sich verschieden verhalten, als ihre Umfänge verschieden sind* (*Nachlass I*, pp. 128, cf HUSS, pp. 319–320).

²⁰ *Man könnte so leicht dahin kommen, den Begriffsumfang für die Bedeutung des Begriffswortes auszugeben; aber hierbei würde man übersehen, dass Begriffsumfänge Gegenstände nicht Begriffe sind*, *Nachlass I*, p. 129. Frege contents himself with the remark: *immerhin ist ein Kern Wahrheit darin enthalten*, *ibid.*

Bedeutung of singular terms) are suitable to perform the duties of *Bedeutung*-importance of singular terms (*solution of problem 3*).

The significance of the biconditional $r \leftrightarrow q$ relative to the conjecture V is obvious; it assures Frege that the truth-values are available exactly when they are needed, if Frege wants them to play the role of *Bedeutung* of sentences (*solution of problem 4*). Naturally, $r \leftrightarrow q$ does not yet secure the suitability of the truth-values; it remains to be established that the truth-value of a sentence $s =$ the truth-value of s' iff s and s' are interchangeable *salva veritate*. While from right to left there is no problem, the left-to-right part is far from evident, especially if claimed with respect to the indefinite totality of ordinary language. This is what keeps Frege from asserting (at least at the beginning) the identity of truth-value and *Bedeutung* of sentences. Frege has to prove the left-to-right part of the suitability condition. The attempt to do this makes up the largest portion of SUB (*solution of problem 5*).

The just mentioned left-to-right conditional is the consequent of the conditional used by Frege in texts D for a second test of his conjecture; let us abbreviate it by " T ". A proof of T , or an argument proving that T is highly plausible (the latter is what Frege thinks of his own achievement towards the end of SUB) has of course a much weaker effect within the second test of the conjecture planned by Frege in texts D than within the looking-around method. Within the second test, the truth of T does not establish the truth of the conjecture, but only confirms the conjecture to the extent that something logically implied by the conjecture is shown to be true. Within the looking-around program, the truth of T is sufficient to establish the suitability of the truth-values (suitability relative to principle B) for being *Bedeutung* of sentences. It might be counted as an objection against our interpretation of Frege in the sense of the looking-around method the fact that he 'wastes' the power of T (at least in the paper SUB) by using T within the second test rather than for a proof of the suitability of the truth-values. It must be admitted that there is an oddity here. I do not think, however, that this is sufficient to disprove that Frege proceeded according to the looking-around method. By the time of writing SUB, it was only a few years that he had first referred to the method in GRL. He was still breaking new ground; the looking-around method could not be as clear to him as it was to CARNAP in *Meaning and necessity* (1956).

The distinction of two *Bedeutungen* in conjunction with the derivation

of BP_1 from B leads us to question the previously asserted logical truth of the conditional $BP_1 \wedge V \rightarrow RS$ (from text C1). BP_1 says that expressions of equal *Bedeutung* may be interchanged without touching the *Bedeutung* of the whole. Now the phrase 'equal *Bedeutung*' has to be read here in the sense of *Bedeutung*-importance, because of our interpretation of BP_1 as derived from B. But the principle RS seems to refer to equal *Bedeutung* in the sense of singular terms standing for, denoting the same object, which is semantic *Bedeutung*. If BP_1 is read with *Bedeutung*-importance and RS with semantic *Bedeutung*, the conditional $BP_1 \wedge V \rightarrow RS$ seems to be no longer logically true. The difficulty, however, is removed by taking advantage of the 'equivalence' of semantic *Bedeutung* and *Bedeutung*-importance in the case of singular terms, which constitute the category of expressions intended by Frege in text C1 (as opposed to D1-D3, where the substituted expressions are sentences).

After basing the interpretation of Frege's notion of *Bedeutung* upon a distinction of two meanings of the word, the issue must be faced of why Frege never mentions the existence of these two meanings. I would like to make the following comments on this.

(a) It is hardly believable that Frege has not been aware of the ambiguity of the German word as used by him for example in the *Begriffschrift*. Consider the following remarks, in a letter to Russell (who had previously complained about the awkwardness of sentences denoting truth-values 'just like' names denote objects, *Nachlass II*, p. 233): *jeder wahre Satz durch jeden wahren Satz unbeschadet der Wahrheit ersetzt werden kann, und ebenso jeder falsche durch jeden falschen. Und damit ist gesagt, dass alle wahren Sätze dasselbe bedeuten oder bezeichnen und ebenso alle falschen Sätze* (*Nachlass II*, p. 247). Undoubtedly 'bedeuten' is here of the importance-for-truth type: Frege cannot have failed to realize that such a 'bedeuten' is not the same as 'bezeichnen'.

(b) Frege may have aimed at fusing the two meanings into one, as revealed by the just quoted text: *bedeuten oder bezeichnen*.

(c) Frege's silence with regard to the two meanings is not really surprising. Frege's theory of *Bedeutung* does not contain, in fact, any *actual* discrepancy between the two meanings. In the case of singular terms there is coincidence: the semantic *Bedeutung* may even play the role of *Bedeutung*-importance. In the case of sentences Frege really considers only one of the two meanings, the *Bedeutung*-importance, so that no discrepancy

can arise. In the case of predicates the situation is more involved. Predicates *bedeuten* concepts,²¹ predicates also *bedeuten* extensions, as we saw, but concepts and extensions are not the same thing for Frege.²² The second 'bedeuten' is surely of the importance type. If the first 'bedeuten' is of the importance type too, we have two kinds of *Bedeutung*-importance and a second, alternative version of principle B is needed. If the first 'bedeuten' is semantical, then we have a conflict between the two meanings of *Bedeutung*. Neither situation, however, ultimately becomes fully real for Frege, because of his peculiar doctrine ruling out the possibility of talking of 'the' *Bedeutung* of predicates.

4. Conclusion

While solving the problems raised by the analysis of the texts, the notion of *Bedeutung* in the proposed interpretation turns out to be disappointing. *What* is the *Bedeutung* of an expression? The answer is: *anything, any entity* that is 'suitable' relative to principle B.²³ Indefinitely many entities may be suitable to play the role of *Bedeutung* of an expression. Frege is aware of the possibility of indefinitely many choices in connection with *Wertverlauf* (GRG I, § 10). Curiously, with regard to *Bedeutung* of sentences he seems to expect that the truth-value of the sentence is the unique suitable candidate (text C3 as well as later writings²⁴). But just the opposite of the truth-value of the sentence would be equally suitable. Now, given that there are many candidates, *why should* one in particular be chosen rather than any other? No sufficient reasons *can* be offered here by the looking-around method. At best, reasons of 'convenience' or 'beauty' may be mentioned (QUINE, 1963, p. 152). This is the essential defect of the looking-around method, hence of Frege's notion of *Bedeutung* in the proposed interpretation. It might be replied that from the point of view of the method no particular entity matters *per se* but

²¹ Ein Begriffswort bedeutet einen Begriff (Nachlass I, p. 128); der Begriff ist Bedeutung eines grammatischen Prädikats (BGGE, p. 193, footnote); diese Worte bedeuten einen Begriff (BGGE, p. 194); das Wort 'Planet'... bezeichnet einen Begriff (UGG, p. 308).

²² Ich meine hiermit nicht, dass Begriff und Begriffsumfang dasselbe sind (HUSS, p. 320).

²³ Cf. QUINE (1963), p. 209: "A cardinal number simpliciter is anything that is \bar{x} for some x " See also p. 152.

²⁴ Nachlass I, p. 211: Das einzige aber...; Nachlass II, p. 240: was kann das Anderes sein, als der Wahrheitswert?

only to the extent that it complies with principle B. This, however, only shows that the method badly needs being reconstructed in terms of abstraction, as originally suggested in the *Begriffsschrift*.

References

- ANGELELLI, I., 1967, *Studies on G. Frege and traditional philosophy* (Reidel)
- ANGELELLI, I., 1976, *The substitutivity of identicals in the history of logic*, in: *Studien zu Frege*, ed. M. Schirn (Frommann)
- ANGELELLI, I., 1979, *Abstraction, looking-around and semantics*, *Studia Leibnitiana*, Sonderheft 8
- CARNAP, R., 1956, *Meaning and necessity* (Univ. of Chicago Press)
- FREGE, G., BG, *Begriffsschrift* (Halle 1879)
- FREGE, G., GRL, *Grundlagen der Arithmetik* (Breslau 1884)
- FREGE, G., KS, *Kleine Schriften* (Darmstadt 1967)
- FREGE, G., FUB, *Funktion und Begriff*, in KS
- FREGE, G., SUB, *Über Sinn und Bedeutung*, in KS
- FREGE, G., BGGE, *Über Begriff und Gegenstand*, in KS
- FREGE, G., HUSS, review of *Husserl's Philosophie der Arithmetik*, in KS
- FREGE, G., UGG, *Über die Grundlagen der Geometrie* I, II, III, in KS
- FREGE, G., GED, *Der Gedanke*, in KS
- FREGE, G., GEFÜGE, *Gedankengefüge*, in KS
- FREGE, G., GRG I, *Grundgesetze der Arithmetik*, I (Jena 1893)
- FREGE, G., GRG II, *Grundgesetze der Arithmetik*, II (Jena 1903)
- FREGE, G., Nachlass I, *Nachgelassener Schriften* (Hamburg 1969)
- FREGE, G., Nachlass II, *Wissenschaftlicher Briefwechsel* (Hamburg, 1967)
- LEIBNIZ, G. W., *Philosophische Schriften*, ed. Gerhardt, vol. VII
- QUINE, W. v. O., 1963, *Set theory and its logic* (Harvard)
- THIEL, Ch., 1965, *Sinn und Bedeutung in der Logik Gottlob Freges* (Meisenheim am Glan)
- TUGENDHAT, E., 1970, *The meaning of 'Bedeutung' in Frege*, Analysis

FROM LEIBNIZ TO FREGE: MATHEMATICAL LOGIC BETWEEN 1679 AND 1879

CHRISTIAN THIEL

Aachen, F.R.G.

Celebrating the centenary of a book is not unusual at all, but it has problems all its own. If the book is rather well known and is a work of high standing, as we may fairly say of Frege's *Begriffsschrift*, then it is as difficult to avoid platitudes as to escape from the temptation to snatch at novel and unthought-of aspects of the work, its background, or its after-effects. I plead guilty to yielding to a similar weakness the day I had to announce the title of my address, a title the pretentiousness of which had been clear to me all along but which grew into a torment in the course of my subsequent studies. But let me jump into the matter by pointing out to what extent it is yet a rewarding enterprise to review the two hundred years of formal logic from 1679 to 1879, even if mathematical logic in the sense of van Heijenoort's *Source Book* (from which the title of this address obviously was purloined), did not spring into life with Leibniz, nor with Boole, as I have been assured by many of my British colleagues, but only with Frege just a hundred years ago.*

Paul Lorenzen, a scholar who will not easily be suspected of exaggerated flattery concerning Platonist logicians, has called Frege's *Begriffsschrift* "a logical masterpiece comparable in originality and import only to Aristotle's *Analytics*"¹ and William and Martha Kneale in their exposition of Frege's doctrines, have come to the opinion that "it is not unfair either to his predecessors or to his successors to say that 1879 is the most im-

* I gratefully acknowledge my debt to Dr. Mark Kulstad for valuable suggestions and criticism of an earlier draft.

¹ LORENZEN, P., 1960, *Die Entstehung der exakten Wissenschaften* (Berlin/Göttingen/Heidelberg), p. 156.

portant date in the history of the subject.² If I do not completely misjudge these and many similar statements, it is not only the singular magnitude of Frege's logical achievements that they refer to; it is also the bewildering impression that Frege created his logic, as it were, *ex nihilo*. And indeed, as to one of his most remarkable achievements, quantificational logic, it would not be appropriate to say that Frege worked in ignorance of previous contributions to the subject, since there was actually nothing he could possibly have ignored; so that, with van Heijenoort, "one cannot but marvel at seeing quantification theory suddenly coming full-grown into the world".³ To sum up, whereas on other logical and philosophical topics, Frege was in fact ignorant of earlier work—as were his contemporaries, and as are still many of ours, in spite of Angelelli's estimable *Studies on Gottlob Frege and Traditional Philosophy*,⁴—regarding the main content of his *Begriffsschrift*, Frege was a genuine pioneer and a finisher at the same time.

Frege himself did not think that he was a pioneer only, but a continuator as well: in the preface to the *Begriffsschrift* he explicitly refers to Leibniz's ideas on a calculus of logic and a *characteristica universalis*. Except for an incidental reference to Aristotle this is the only reference of this kind in the whole book. It was only as a reaction to Schröder's unfavourable review of the *Begriffsschrift* that he released some information about his previous study of logical systems, but there is no reason for suspecting that he read Boole, McColl, Robert Grassmann, and Schröder only after the latter's critique.^{5,6} Personally, I am a little puzzled by Frege's remark in the same preface that reads: "Quite alien to my mind have been those endeavours to create an artificial similarity [i.e. between logic and arithmetic, Th.] by regarding a concept as the sum of its marks [Merkmale, Th.]" (Bs., p. IV). This formulation has been connected with Schröder's calculus of domains by Jourdain and by Wilma Papst (for which there is some support from the posthumously published

² KNEALE, W. and M., 1962, *The Development of Logic* (Oxford), p. 51.

³ VAN HEIJENOORT, J., 1967, Introduction to *Begriffsschrift*, in: From Frege to Gödel. A source book in mathematical logic, 1879–1931, ed. J. van Heijenoort (Cambridge, Mass.) p. 3b.

⁴ ANGELELLI, I., 1967, *Studies on Gottlob Frege and Traditional Philosophy* (Dordrecht).

⁵ SCHRÖDER, E., 1881, *Review of Begriffsschrift*, *Zeitschrift für Mathematik und Physik*, vol. 25, pp. 81–94.

⁶ FREGE, G., 1883. *Ueber den Zweck der Begriffsschrift*, *Jenaische Zeitschrift für Naturwissenschaft* vol. 16 (1883), Supplement, pp. 1–10.

early manuscripts), but taken literally it would refer to intensional interpretations of logical calculi, manifesting Frege's familiarity with this subject of contemporary controversies (a reading supported by Frege's later avowal in favour of extensional logic in opposition to a logic of content).

More mysterious than this detail for devoted historians is the fact that, apart from the statement that Frege's starting point for his train of thought leading to the *Begriffsschrift* was arithmetic (Bs., p. VIII), and a tiny remark (Bs., p. 4) disclosing that he had once experimented with another system, there is absolutely no trace of any "prehistory", neither as to previous occupation with logic, nor as to the development of that peculiar notational system which Frege remained the only one to use. Not a single one of the seventeen lectures and reports that Frege gave to the Mathematical Society of Jena between 1866 and 1881 seems to have anything to do with logic—and the seven of which summaries in Frege's own hand have been preserved provably do not.

Taking Frege's pioneer work in logic and its impact on the present state of the art to be sufficiently investigated, let us ask what had been going on in logic since Leibniz—some of whose "germs of thought" Frege decided to foster—and let us ask why logic between Leibniz and Frege had practically no influence at all on the latter. For the regrettably short and incomplete survey that can be given here, it will suffice to outline the most important directions and to indicate particular achievements. Roughly, I will touch upon Leibniz, on the intensional logic of the so-called Leibniz School, on Saccheri, on the algebra of logic with its treatment of classes, propositions, and relations, and finally upon Frege again. I will, however, abstain from a chronological enumeration of works and achievements, most of which will be known to this audience, at least regarding the algebra of logic. Instead, I will use, or if you like, abuse this presentation for arguing a case that I believe deserves consideration and support at an occasion like this meeting to commemorate what van Heijenoort has called "perhaps the most important single work ever written in logic"⁷. What I wish to argue for is nothing less than a reconsideration or even re-assessment of the historiography of logic. And while I do not have complaints about the presentation of Frege in our texts on history of logic, I certainly do have complaints about many other cases, some of which I will mention later.

⁷ VAN HEIJENOORT, J., loc. cit. (see note 3), p. 3.

It seems that Leibniz was the first scholar in possession of a clear conception of a calculus as a set of rules for performing operations of a strictly determined kind on graphical patterns, whether on strings of letters taken from some alphabet, or on geometrical diagrams. It is well known that he made at least three differentiated attempts at establishing a calculus of logic. In his first attempts, he tries to turn elementary number theory, to wit, the theory of divisibility, to advantage in the field of Aristotelian categorical syllogistic. After having failed with the assignment of prime numbers to elementary concepts and composite numbers to complex concepts, this leaving out of account what has been called "negative concepts", Leibniz associates pairs of numbers with the subject and the predicates of propositions:

$$\begin{aligned} S &\leftrightarrow \langle s_1, s_2 \rangle, \\ P &\leftrightarrow \langle p_1, p_2 \rangle, \end{aligned}$$

in order to express the traditional standard forms of syllogistic in the following form:

$$\begin{aligned} SaP &\leftrightarrow p_1 | s_1 \wedge p_2 | s_2, \\ SiP &\leftrightarrow (p_1, s_2) = 1 \wedge (s_1, p_2) = 1, \\ SoP &\leftrightarrow \neg SaP, \\ Sep &\leftrightarrow \neg SiP \end{aligned}$$

(where (a, b) is the greatest common divisor of a and b). Rewriting the premisses of a syllogism arithmetically in the manner indicated, one can check whether the correlate of the conclusion may be inferred also in arithmetic. Until quite recently, it was almost generally agreed upon that this calculus fails, too, and that Leibniz did not follow it up just because of this insight. Łukasiewicz in his book on Aristotelian syllogistic is the only one I know of who explicitly states that Leibniz's calculus no. 1 is a correct arithmetical interpretation of categorical syllogistic. Whether it is or not depends on the definition of validity. In a paper read at the Third International Leibniz Congress here in Hanover in 1977, I have tried to show that Leibniz first had an incorrect definition of validity and tended to give up his arithmetical interpretation, but that he himself found the correct definition immediately afterwards, and that with this definition the calculus is entirely correct.⁸

⁸ THIEL, Ch., *Leibnizens Definition der logischen Allgemeingültigkeit und der "arithmetische Kalkül"*, in: *Theoria cum praxi. Proceedings of the III. International Leibniz Congress, Hanover 1977* (Wiesbaden 1980), vol. 3, pp. 14–22.

So we do *not* know why Leibniz abandoned the arithmetical calculus, but we can be sure the reason was *not* its alleged failure in the domain of syllogistic. The reason may have been very trivial: Leibniz was not occupied with logic all the time; between 1675 and 1684 he made decisive progress with his infinitesimal calculus and may have been taken up with problems in this field. On the other hand, it is quite clear that even an operative arithmetical calculus for syllogistic was far too narrow for Leibniz's conception of logic, which tended to a *general* calculus of concepts, extensional and intensional, and was not restricted to the traditional standard forms like "Every *A* is *B*", and which moreover was conceived to take up and extend *Joachim Jungius'* treatment of non-syllogistic inferences in the *Logica Hamburgensis* (1638, a work highly esteemed by Leibniz), which means, to work in the direction of a logic of relations, and finally to include a logic of propositions by extending Leibniz's general calculus *de continente et contento* to relations of entailment between propositions. It is to *both* extensions that Leibniz's statement in the *Nouveaux Essais* refers, saying that, "there exist sound asyllogistic inferences which it would be impossible to prove by any syllogism".⁹ The result, in Leibniz's second and third attempt at a calculus of logic (1686 and 1690), are some fragmentary lattice-theoretical systems, admitting of various interpretations, extensional, intensional, geometrical, and others, "Every *A* is *B*" being interpreted, e.g., intensionally as "*A* = *AB*" in 1686, and as "*A* = *A+B*" in 1690, when Leibniz had a calculus with +, -, and =, ≠, and <.

These calculi have been expounded so often and in detail that I will abstain from another description of them, and just mention a so far uncorroborated conjecture of mine concerning Leibniz's change to the so-called "algebraic calculi" that I have called "lattice-theoretical" in the preceding sentence. In more elaborate books on the history of logic, we find James, i.e. Jacob Bernoulli mentioned with his book, *The Parallelism between Logical and Algebraic Arguments*¹⁰, and we are told, e.g. in Styazhkin's *History of Mathematical Logic from Leibniz to Peano*¹¹, that Bernoulli, "following Leibniz ... only noted the analogy that exists between the laws of formal logic and the methods of elementary algebra"

⁹ NE IV, XVII § 4; see also II, § 9.

¹⁰ BERNOULLI, JAC. I, *Parallelismus ratiocinii logici et algebraici*, Basel 1685; also in: *Opera*, ed. G. Cramer, I, Geneva 1744, repr. Brussels 1968.

¹¹ STYAZHKIN, N. I., 1969, *History of Mathematical Logic from Leibniz to Peano* (Cambridge, Mass./London) (Russian original 1964).

(p. 95). As I do not know of any algebraically oriented logical work done by Leibniz previous to 1685—leaving out of account the combinatorial investigations into syllogistics in the *Dissertatio*¹²—I would suspect that Leibniz's experiments with algebraic calculi were stimulated by Bernoulli's work in spite of the fact that the latter was composed rather sketchily and lacked profundity, so that Leibniz certainly could not have profited from this work to the same extent as later from Jacob Bernoulli's insights into probability.¹³ And indeed, Leibniz's conception had become so broad as to render the exploitation of algebraical methods more difficult: In his letter to Gabriel Wagner on the utility of the art of reasoning or logic (1696) he states: “I have come to the opinion that even algebra borrows its advantages from a much higher art, to wit, from the true logic”.¹⁴

The question whether Leibniz deserves an eminent place in the history of mathematical logic is not sacrilegious. The answer depends not only on our definition of “mathematical logic” but also on our conceptions of history and historiography. If history is what actually happened and had a significant effect, then Leibniz has to be omitted from the history of logic in his lifetime, for the writings referred to as containing his great and fruitful ideas in logic were completely unknown to his contemporaries. Part of them did not appear in print until Raspe's edition of 1765, and most of them appeared only at the beginning of our century when Couturat made them accessible in 1901 and 1903, such that in Leibniz's lifetime it is mainly his remarks in letters to contemporary scholars that could have had any traceable influence. He who conceives of the history of logic as the *development* of logical ideas, however, would be entitled to assign a rather minor place to Leibniz when talking about the logic of the *Neuzeit*, and I may for similar reasons be forgiven my decision to omit the logical work of Bolzano and of Robert Grassmann from this paper as they did not really influence the development of logic within the period we are concerned with.

Before passing to the so-called *Leibniz School* in logic, I should say a few words on *Gerolamo Saccheri* (1667–1733) who is mainly being remembered for his *Euclides ab omni naevo vindicatus*, i.e. *Euclid Cleared of Every Flaw*, published in Milan in 1733, and containing a good portion

¹² LEIBNIZ, G. W., 1666, *Dissertatio de Arte Combinatoria* (Leipzig).

¹³ BERNOULLI, JAC., 1713, *Ars Conjectandi*, Basel, repr. Brussels 1968.

¹⁴ LEIBNIZ, G. W., 1696, *Schreiben an Gabriel Wagner. Vom Nutzen der Vernunftkunst oder Logik*. Erdm. 418 ff., quotation on p. 424 b.

of unstrived-for development of theorems of a non-Euclidean geometry. Less known is Saccheri's *Logica Demonstrativa*, which saw three editions, an anonymous one in 1697, the others in 1701 and in 1735.¹⁵

This work deserves a place in the history of logic not only for Saccheri's interest in a logical analysis and description of mathematical proof techniques in geometry, but also because of two contributions presented in its eleventh chapter. Saccheri here develops a "via nobilior", i.e. a "more noble way" than previously followed, in logic generally as well as earlier in his book. His first idea is that of proving a theorem by deriving it from its own contradictory.¹⁶ According as the theorem is affirmative or negative, this means proceeding by

$$(\neg a \rightarrow a) \rightarrow a \quad \text{or} \quad (a \rightarrow \neg a) \rightarrow \neg a,$$

or rather,

$$\frac{\neg a \rightarrow a}{a} \quad \text{and} \quad \frac{a \rightarrow \neg a}{\neg a},$$

i.e., as rules of inference. Saccheri's goal was a consistency proof for (or rather, a proof of the truth of) a postulate system, say for geometry, by applying this "principle of necessary truth", i.e. deriving each of the postulates from its own contradictory. This method has been taken up in our century by Josiah Royce¹⁷. Indeed,

$$\frac{\neg a \rightarrow a, \neg b \rightarrow b, \dots}{\neg \neg(a \wedge b \wedge \dots)}$$

¹⁵ AUGUSTAE TAURINORUM (Torino) 1697; Ticini Regii (Pavia) 1701 (Heinrich Scholz's correction of Vailati's date "1701" to "1702" in his *Abriss der Geschichte der Logik*, Freiburg/München ^a1959, ^b1967, p. 37, is itself mistaken); Augustae Ubiorum (Köln) 1735.

¹⁶ See p. 80 of the 1697 edition: "Sumam contradictorium propositionum demonstrandarum, ex eoque ostensiùè, ac directè propositionem eliciam" (p. 82 of the 1701 and p. 130 of the 1735 edition). As will be seen, my interpretation diverges from Angelelli's and Hoorman's who identify this principle with Clavius' Law, and the *via negativa* with indirect proof without distinguishing *reductio ad absurdum* from $a \rightarrow \neg a \leftarrow \neg a$. Saccheri, in the Scholium to chapter 11, presents the *via negativa* as a *deductio ad impossibile* as is further corroborated by his example, "v. g. quod, si modus AA concluderet in secunda figura, omnis, vel aliquis syllogismus AA esset syllogismus EA." (p. 89 of the 1697 edition).

¹⁷ JOSIAH ROYCE, *Prinzipien der Logik*. In: *Encyclopädie der philosophischen Wissenschaften*, in Verbindung mit Wilhelm Windelband herausgegeben von Arnold Ruge. Erster Band: *Logik*. Tübingen 1912, pp. 61–136. The original English version appeared in the first volume of the (unfinished) English edition: *The Principles of Logic*. In: Sir Henry Jones (ed.), *Encyclopaedia of the Philosophical Sciences*, vol. I: Logic (London 1913), pp. 67–135. A separate edition was published in New York in 1961.

is even constructively valid. This method was put to use in Saccheri's *Euclides*, and earlier in his *Logica Demonstrativa* in connection with the *via nobilior* for a justification of the syllogisms by a sort of self-application which I do not have the time here to expound but on which we have valuable commentaries by Angelelli 1975, Hamblin 1975, and Hoorman 1976.¹⁸

Whereas Saccheri worked independently of Leibniz's logical experiments, the writers of the so-called Leibniz School were influenced by Leibniz although part of their work precedes Raspe's edition of 1765. For example, Johann Andreas von Segner's *Specimen Logicae* appeared in Jena in 1740, Georg Johann von Holland's correspondence with Lambert started previous to 1764, the year in which the first two volumes of Johann Heinrich Lambert's *Neues Organon* appeared in print. All these writers, stimulating each other by correspondence and small treatises, worked along Leibnizian lines, but restricting themselves (with few exceptions) to the *intensional* interpretation of their calculi. Lambert makes use of a quantification of the predicate—a device concerning which Hamilton and De Morgan claimed the priority of invention some sixty years later—and distinguishes two cases of each standard form, e.g. in the universal affirmative proposition "All *A* is *B*", Case I: $A = B$ where the converse is also universal, and Case II: $A < B$ where the converse is particular. In other cases, a standard form had to be represented by two formulae, and the calculi became very complicated, too complicated for Ploucquet, Lambert, Holland and others to handle in a satisfactory manner. John Venn in his *Symbolic Logic* has attributed this failure to the intensional standpoint which he also blamed for the incompleteness and inutility of Leibniz's endeavours. This has become the standard value judgment on the Leibniz School, including *Castillon*, whose calculus was considered to be the most consequential. Couturat has strongly emphasized this point of view with regard to Leibniz. C. I. Lewis stated it plainly: "This movement produced nothing directly which belongs to the history of symbolic logic. ... The record of symbolic logic on the continent is a record of failure, in England,

¹⁸ ANGELELLI, I., 1975, *On Saccheri's use of the "Consequentia Mirabilis"*, Akten des II. Intern. Leibniz-Kongresses 1972, Bd. IV, (Wiesbaden 1975), pp. 19–26; C. L. HAMBLIN, *Saccherian arguments and the self-application of logic*, Australasian Journal of Philosophy, vol. 53, pp. 157–160; CYRIL F. A. HOORMAN, Jr., *A further examination of Saccheri's use of the "Consequentia Mirabilis"*, Notre Dame Journal of Formal Logic vol. 17 (1976), pp. 239–247.

a record of success. The continental students habitually emphasized intension, the English, extension”¹⁹.

Meanwhile, formal logic on the continent had suffered a serious set-back at the hands of Kantian transcendental logic which, though in no way incompatible with formal logic, detracted, at least in Germany, interest from formal logic which, moreover, was treated very contemptuously by Kant whose own lectures on logic might have been quite suitable to support this attitude, but who had ridiculed Leibniz’s idea of a logical calculus and a universal characteristic already in his *Nova Dilucidatio*.²⁰ Hegel’s judgment, and Lotze’s in 1841, were in no way more favourable, and in spite of my respect for investigations into the very *foundations* of formal logic as we find them in most textbooks of “logic” during the last century, I cannot overlook the fact that the intellectual climate on the continent had indeed become suffocative for symbolic logic by the middle of the 19th century.

Typical of this is the fate of Moritz Wilhelm Drobisch’s logic in its original form, entitled *A New Exposition of Logic, According to Its Simplest Relations, with a logico-mathematical Appendix*²¹. The appendix must be mentioned here since in the first edition, there is a paragraph with the title, “algebraic construction of the simplest forms of judgment and a derivation of inferences founded thereupon”²², where Drobisch develops an *extensional* calculus of classes and elementary judgments, improving the intensional systems of Ploucquet, Lambert, Jacob Bernoulli, and Gergonne, all of whom he explicitly mentions.

Unfortunately, Drobisch felt obliged to withdraw this valuable paragraph in the second and later editions because of a severe criticism put forward by Friedrich Adolf Trendelenburg, who in his *Logical Investigations*²³ criticized formal logic as taught by Leibniz, Twesten, Drobisch and others

¹⁹ LEWIS, C. I., 1918, *A survey of symbolic logic* (Berkeley, abridged New York 1960), p. 36 f.

²⁰ KANT, I., *Principiorum primorum cognitionis metaphysicae nova dilucidatio* (&c.), 1755, Akad. I. (Berlin 1902/10), pp. 385–416 (esp. p. 390).

²¹ DROBISCH, M. W., *Neue Darstellung der Logik nach ihren einfachsten Verhältnissen. Nebst einem logisch-mathematischen Anhang* (Leipzig 1836); *Neue Darstellung der Logik nach ihren einfachsten Verhältnissen, mit Rücksicht auf Mathematik und Naturwissenschaft*, Zweite, völlig umgearbeitete Auflage (Leipzig 1851); Dritte neu bearbeitete Auflage (Leipzig 1863); Vierte verbesserte Auflage (Leipzig 1875); Fünfte Auflage (Leipzig 1887).

²² op. cit. (see note 21), *Algebraische Construction der einfachsten Urtheilsformen und darauf gegründete Ableitung der Schlüsse*, pp. 131–136 of the 1836 edition.

²³ TREDELENBURG, F. A., *Logische Untersuchungen* (Berlin 1840, Leipzig 1862).

for its allegedly unjustified treatment of concepts by arithmetical analogies. It is a curious fact that the same Trendelenburg, who claimed that Leibniz did not conceive logic as a formal discipline at all, published in 1856 an essay, *On Leibniz's Project of a Universal Characteristic*²⁴, in which he also sketched the prehistory of this project since Raymundus Lullus, thereby drawing attention to the universal language movement, and contributing essentially to the revival of interest in Leibniz's logic. Incidentally, this essay of Trendelenburg of 1856 is the earliest text in which I have been able to find the German expression *Begriffsschrift*, although used in a very general sense for any notation that brings the formation of a sign in touch with the content of the concept it denotes, as Trendelenburg finds it in numerals and other implements of science.

The so-called *algebra of logic*, crowned with the names of De Morgan, Boole, Peirce, and Schröder, has been so thoroughly investigated and expounded in the literature that it can be taken to be the best known period in the history of formal logic. Let me just remind you of De Morgan's foundation of the theory of relations, surpassed only by Peirce's work in the subject, whose calculus of relations could already profit from Boole's calculus of classes. Boole had started with a successful attempt at an axiomatized Aristotelian syllogistic, in order to defend one of De Morgan's claims in his quarrel with Hamilton. The result was *The Mathematical Analysis of Logic, being an Essay towards a Calculus of Deductive Reasoning* (Cambridge and London 1847), the exposition of the intended system being given in *An Investigation of the Laws of Thought, on which are founded the Mathematical Theories of Logic and Probability* (London and Cambridge, 1854).

Boole was either ignorant or intentionally negligent of the logical work of his immediate predecessors, and he did not devote much thought to the problems of extension and intension, of existential import, and of empty classes, problems that had hampered previous logicians and were to return later, but of which C. I. Lewis says rightly: "It is well that, with Boole, they are given a vacation long enough to get the subject started in terms of a simple and general procedure" (Survey, p. 51). A peculiarity

²⁴ TRENDELENBURG, F. A., *Über Leibnizens Entwurf einer allgemeinen Charakteristik* (Vorgetragen zur Feier des Leibnitztages 1856). Philosophische Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin. Aus dem Jahre 1856 (Berlin 1857), pp. 37–69 (also published separately); reprinted in F. A. T., *Historische Beiträge zur Philosophie. Dritter Band. Vermischte Abhandlungen* (Berlin 1867), pp. 1–47. The expression "Begriffsschrift" on p. 39 of the 1857, p. 4 of the 1867 printing.

of Boole's, which he shares with Frege, is his conviction of the inter-relation of logic, language, and mathematics, a thought that impressed itself on many of his logical conceptions, perhaps even on the parallelism claimed by Boole to hold between the algebra of classes and that of propositions or of 0 and 1, later adjusted by Peirce who pointed out the special status of the fact that if $x \neq 0$, then $x = 1$.

Let me return to Frege by relating his *Begriffsschrift* to the algebra of logic, following his admirably clear statements in his paper, *On the Purpose of the "Begriffsschrift"* (1882), written as an address to the Jena Society for Medicine and Natural Sciences to defend himself against Schröder's reproach that Frege was, firstly, ignorant of earlier work and that secondly, his *Begriffsschrift* marked a regress in comparison with the algebra of logic. Frege points out that his purpose was *not* the presentation of abstract logic by formulae, but the creation of an instrument capable of expressing a content in a more precise and more perspicuous way than could be done by natural language. He finds fault with Boole's classification of propositions into primary and secondary ones, this leaving out of account existential propositions, and misses in Boole's system a notation for individuals (as different from singletons), and for the falling of an individual under a concept. Moreover, granting the possibility of interpreting Boole's calculus as dealing with classes on the one hand, and as dealing with propositions on the other hand, Frege does not find a bridge between the two, a bridge for passing from a proposition to the concepts contained in it, and vice versa. And as to the notation, this latter problem makes "0" and "1" ambiguous in arithmetical contexts where "0" and "1" already have their ordinary numerical meaning. To sum up, Frege thinks that Schröder was mistaken in trying to compare two systems of logic which depart from different starting points and aim at the fulfilment of different purposes. As manuscripts from the *Nachlass* show, it was quite clear to Frege that, in spite of the term "*Begriffsschrift*", which in this respect is rather ill-chosen, he did not start from concepts in order to combine them into propositions, but from propositions, decomposing them into function and arguments. Whereas in other respects he views himself as in the tradition of Leibniz, he is quite aware of the fact that here his system is opposed to that tradition.

In my opinion, it will not be necessary to go into the system of the *Begriffsschrift* here. Quantificational logic in Frege's setting is close enough to its modern formulation to be considered well known itself, and I will mention only a point that is usually misunderstood. Introducing a notation

for quantification, i.e. the universal quantifier and bound variables, Frege explicitly says that in “ $\Phi(A)$ ”, “ Φ ” may be considered as the argument of a function, too, and may then be replaced by a German letter in a quantified proposition. It is, therefore, incorrect to say that Frege established only first-order logic in the *Begriffsschrift*, and that he illegitimately substituted bound functional letters for bound individual letters in some places of the third part of the *Begriffsschrift* where the concepts of sequence and successor are being defined. In the *Begriffsschrift*, the “ a ” of “ $\underline{\underline{a}} \Phi(a)$ ” is typically ambiguous.²⁵

To many of you, what I have said in this paper will be proof enough for the delicacy of the problem of how to write or even to approach the history of logic. There are, first and almost trivially, two directions of approach. One may include everything that seems relevant from the view-point of the historiographer's time, or one may, on the other hand, try to do justice to historical authors by taking seriously their contemporary problems, and even by trying to gain a better understanding of them by looking for their origins.

Secondly, we have to make a decision whether we want to write a history of causes and effects, a history of stimuli and influences, or a history of aims and reasons, whether we want to investigate *Wirkungsgeschichte* in the traditional sense which will not be changed in principle by incorporating social and economical aspects and developments (desirable as this would be for the history of formal logic as it has been done for mathematics to some extent at least), or what Jürgen Mittelstrass has called *Gründegeschichte*. My vote is for the latter, a history of means and ends, aims and reasons, of needs, goals, and purposeful actions. This does not strip *Wirkungsgeschichte* of its status as an indispensable basis of *Gründegeschichte*.

²⁵ This is to correct van Heijenoort's remark on p. 3 of his introduction to Frege's *Begriffsschrift* quoted in note 3 above. The correctness of Frege's inferences has also been pointed out by Terrell Ward Bynum (*On an Alleged Contradiction Lurking in Frege's Begriffsschrift*, Notre Dame Journal of Formal Logic, vol. 14 (1973), (pp. 285–287) who thinks that while Frege “does not yet have all the machinery or the terminology to precisely spell out the distinction between what he would later call ‘first-level’ and ‘second-level’ functions, he never confuses the two” (p. 285). But closer inspection of Frege's text shows (end of § 10, § 11, i. e. pp. 18–19, and pp. 60–62, p. 68 footnote) that the typical ambiguity is fully intentional. Indeed, in the *Begriffsschrift* expression

“ $\underline{\underline{a}} \Phi(a)$ ” the letter “ a ” stands for “argument”, the letter “ Φ ” for “function”.

geschichte which I would like to be understood as a history of (in our case) logic with particular emphasis on reasons for the various developments, but as well as a history of reasons themselves, their interconnections and changes.

No doubt this is a beautiful challenge and programme, but I am afraid there is some troublesome work to be done before it can be taken up: there are scores of false data and erroneous judgements to be cleared away, and to be replaced by correct and more reliable ones. Let me just mention a small number of instances. Most historians of logic have not been able to see personally all the texts they quote and refer to; they rely on older historians of logic and carry over incorrect data as well as erroneous evaluations from these. How else could it be explained that *Bardili*, *Victorin*, *Twesten* are counted among the early representatives of mathematical logic in many histories of logic although there is practically nothing mathematical in them except variables *A*, *B*, ... and perhaps the plus, the minus and the equality sign? Why is Twesten referred to as a Leibnizian and as a member of the Leibniz School, while being strongly influenced by Kant, the structure of whose logic he took over? Why is *Friedrich von Castillon* (1747–1814), author of *Réflexions sur la logique* (1802) and the *Mémoires sur un nouvel algorithme logique* (1803), mistaken for his father, Jean or Giovanni Francesco de Castillon (1708–1791), practically everywhere, from John Venn via Shearman, C. I. Lewis, A. Church's *Bibliography*, Bocheński's *Formal Logic*, Styazhkin's *History* up to modern authors of 1970?²⁶ Going back to primary sources as far as they are accessible, and searching for items that have seemingly disappeared, seem to me the presently most important tasks to be taken care of before a reliable and informative history of mathematical logic can be written. It would be unfair and incompetent not to mention the work that has so far come closest to the demand stated: *Wilhelm Risse's Die Logik der Neuzeit*, a comprehensive enterprise that has so far grown into two volumes covering the period from 1500 to 1780. But it cannot be overlooked that Risse's is a history of logic in a broad sense, mathematical logic playing a subordinate part except in Leibniz, and even here there is a certain lack of apprehension of more recent research done by others with results that might have illuminated or made accessible obscure or difficult texts and passages. A true re-assessment of the history of the

²⁶ See my paper *Zur Beurteilung der intensionalen Logik bei Leibniz und Castillon*, Akten des II. Intern. Leibniz-Kongresses, Hannover 1972, Bd. IV (Wiesbaden 1975), pp. 27–37.

subject will require much closer cooperation between historians and mathematical logicians; let us hope that it will come about.

Let me come to an end with a short systematical remark that might be relevant for the historical evaluation of Frege's logic. As is well known, Frege developed his *Begriffsschrift* further and presented an intricate and profound system of logic and part of set theory in his *Grundgesetze der Arithmetik*, the first volume of which was published in 1893, the second in 1903. The destiny of this system is widely known: it proved inconsistent by admitting the derivation of the Zermelo-Russell antinomy. But after the wreckage of the *Grundgesetze* by this antinomy, Frege and other logicians came to believe that the system might be saved by modifying the principle of abstraction which had served to supply Frege's courses-of-values. Leśniewski and later Quine have shown that Frege's own modification was insufficient for his purposes, but there remains the simple question whether by seizing the principle of abstraction we got hold of the real trouble-maker.

As to this question, it seems to be little known even among experts that the fuse leading to the Zermelo-Russell antinomy is already hidden in the *Begriffsschrift* in spite of the absence of courses-of-values or extensions of concepts in this system. In what sense this claim is justified will become clear from an analysis of Frege's Appendix to the second volume of the *Grundgesetze*. After presenting a meticulous derivation of the antinomy, and a shorter version using his elementhood symbol " \cap ", Frege blames his fundamental law V for the catastrophe. To be more precise, he blames one direction of it. Splitting (and slightly changing)

$$\vdash (\dot{e}f(\varepsilon) = \dot{a}g(a)) = (\neg \text{---}^{\alpha} f(a) = g(a)) \quad (\text{V})$$

into

$$\vdash \begin{array}{l} F(\dot{e}f(\varepsilon)) = F(\dot{a}g(a)) \\ \quad \text{---}^{\alpha} f(a) = g(a) \end{array} \quad (\text{Va})$$

and

$$\vdash \begin{array}{l} f(a) = g(a) \\ \quad \dot{e}f(\varepsilon) = \dot{a}g(a), \end{array} \quad (\text{Vb})$$

the latter part appears to be essential for Frege's derivation of the antinomy. To raise this above all doubt, Frege decides to make his argumentation independent of courses-of-values altogether by deriving the falsehood of V_b without reference to them. The idea is to consider V_b as a special case of

$$\vdash \boxed{f(a) = g(a)} \\ \vdash M_\beta(\neg f(\beta)) = M_\beta(\neg g(\beta))$$

and to derive the negation of this formula in the system. Checking this derivation, one finds that Frege makes use of the fundamental laws IIa and IIb, of theorems Ig, IIIa, and IIIe, and that he uses substitution, detachment, contraposition, transitivity, and universal generalization as his rules. But substitution and detachment are precisely the rules of the *Begriffsschrift* system, the other rules appealed to are provably admissible, the fundamental law IIa is the *Begriffsschrift* axiom 58, theorems IIIa and IIIe are axioms 52 and 54, while Ig can easily be proved as a theorem of the *Begriffsschrift*, and IIb is understood by Frege to be contained in axiom 58, as indicated above and as may be seen from his § 10 and his procedure in the third part of the text, especially on pp. 60–62.²⁷

In other words, the derivation in the Appendix can be carried out in the *Begriffsschrift* of 1879 as well, the result being a theorem stating that for any second-level function that takes an argument of the second kind (i.e., a one-place first-level function), there are two concepts yielding the same value when taken as arguments of the function although there are objects falling under one of them but not under the other. As this is valid for any second-level function whatsoever, it will also hold for Frege's course-of-values function $\dot{\epsilon}\phi(\varepsilon)$, contrary to its purpose: there are concepts $\Phi(\xi)$ and $\Psi(\xi)$ which yield the same value when taken as arguments, i.e. $\dot{\epsilon}\Phi(\varepsilon) = \dot{\epsilon}\Psi(\alpha)$, although there is at least one object A such that $\Phi(A)$ but not $\Psi(A)$. Hence, $\Phi(\xi)$ and $\Psi(\xi)$ do *not* both extend to the same objects; their extensions are different. As $\dot{\epsilon}\Phi(\varepsilon) = \dot{\epsilon}\Psi(\alpha)$, neither $\dot{\epsilon}\Phi(\varepsilon)$ nor $\dot{\epsilon}\Psi(\alpha)$ can be the extension of the concepts $\Phi(\xi)$ and $\Psi(\xi)$ in the traditional sense. This, of course, obstructs Frege's definition of number as an extension (or course-of-values)—unless some other step or expression in his derivation can be made responsible for the unpleasant

²⁷ Cf. note 25.

result. The problem of identity of sets has returned through the back door as the problem of identity of functions.

I will pursue this question further elsewhere, and restrict myself to the remark that the feature of Frege's *Begriffsschrift* just sketched refutes the wide-spread opinion that his system of 1879 is not only less precise than that of 1893—which is true—but also essentially weaker. And let us not forget, lastly, that it is, in spite of certain deficiencies, the beginning of mathematical logic in the narrower and modern sense, even though mathematical logic a century ago means mathematical logic *before* Peano's axioms of 1889, *before* Hilbert's *Foundations of Geometry* of 1899, *before* Zermelo's axioms of set theory, *before* Brouwer's intuitionistic critique of 1907, *before* *Principia Mathematica*, *before* Hilbert's *Programme*, *before* recursive functions, and *before* Gödel's theorems. Was not Frege's *Begriffsschrift* the memorable first step of a flight of stairs leading to a rapidly expanding world of logic?

RATIONAL ALLOCATION OF RESOURCES TO SCIENTIFIC RESEARCH

PATRICK SUPPES

Stanford University, Stanford, California, U.S.A.

1. Ingredients of decision making

Since World War II, a major research effort has been devoted to the theory of decision making. To a considerable extent the general theory has been closely identified with working out a proper conceptual foundation for mathematical statistics. Evidence of this is the fact that probably the most influential general work in decision making has been L. J. SAVAGE's *Foundations of Statistics* (1954). On the other hand, there is already a large literature concerned with particular areas of application ranging from linear programming models for petroleum refineries to models of accountability for school systems. A good general reference on applied decision theory is RAIFFA and SCHAFLER (1961).

Let us turn now to the basic ingredients of the general theory. First, we have the set S of possible states of nature. Introduction of this set reflects the fact that we do not know the true state and thus are always in the position of making decisions in the face of uncertainty. Second, we have the set C of possible consequences or, as Savage puts it for the most general model, the set of possible future histories of the universe. In practice, of course, we consider always a much reduced model and examine only proximate consequences of a given decision or action. The third ingredient is the set of possible decisions; formally, each decision is a function from the set S of possible states of nature to the set C of possible consequences. Thus, given a decision d and a state s of nature, $d(s) = c$ for some possible consequence c that is a member of the set C . The fourth ingredient is a relation \geq of preference among decisions.

Axioms of rationality are then imposed on these ingredients. For example, almost everyone would agree that a rational decision maker must

have a transitive relation of preference; that is, for any three possible decisions d , f , and g , if $d \geq f$ and $f \geq g$, then $d \geq g$. It is not easy, and perhaps not possible, to have a purely intrinsic criterion of what axioms on preferences among decisions are to be regarded as rational. Fortunately, a simple extrinsic criterion exists and is widely accepted. According to this criterion, we need axioms that are strong enough to enable us to prove that satisfaction of the axioms by a rational decision maker guarantees the following:

- (i) the existence of a unique subjective probability distribution p on the set S of states of nature;
- (ii) the existence of a numerical utility function U , unique up to a positive linear transformation, on the set C of possible consequences;
- '(iii) a choice among decisions that satisfies the rationality criterion of maximizing expected utility, where the expectation is with respect to the probability distribution p , that is, for any two possible decisions d and f ,

$$d \geq f \text{ if and only if } \sum_{s \in S} p(s) U(d(s)) \geq \sum_{s \in S} p(s) U(f(s)).$$

Note that $E(d)$, the expectation of d , is just the sum of products shown above, that is,

$$E(d) = \sum_{s \in S} p(s) U(d(s)),$$

and so the criterion of maximizing expected utility can be expressed in simpler terms as:

$$d \geq f \text{ if and only if } E(d) \geq E(f).$$

Two detailed surveys of a variety of work on the general theory, including some experimental studies, are LUCE and SUPPES (1965) and KRANTZ, LUCE, SUPPES, and TVERSKY (1971, Chap. 8).

A more particular reference is the normative theory of development outlined in Chapter 5 of MARSCHAK, GLENNAN, and SUMMERS (1967), which builds on the general theory of decisions but moves on to a different set of questions than I consider in the remainder of this article. The concern is with optimal policies of development for specific goals rather than with rational allocation of resources for basic research, but, all the same, the ideas developed by Marschak (the author of Chapter 5 of the book) are worth applying to research allocations as well; limitations of time and space have prevented my making the effort in the present context.

2. A schematic example using Pakistani science data

Without attempting to deal realistically with science policy in Pakistan, it may still be useful to build a highly simplified model that uses some Pakistani data on scientific research. For this purpose I use the data for 1966–67 cited by Abdus SALAM (1970). I show in Table 1 the allocations he cites for research in various sectors. We may take the total allocation of 7.39 crores as given, fixed by a higher level of government, and assume that the relevant science policy question is the allocation to each sector.

Table 1
Allocation of Research Funds to Various Sectors

| | Sector | Crores |
|-----|--|--------|
| (1) | Industrial research | 1.92 |
| (2) | Atomic energy research | 1.94 |
| (3) | Agricultural research | 1.80 |
| (4) | Environmental sciences | 0.79 |
| (5) | Medical and family-planning research | 0.29 |
| (6) | Building and roads research | 0.16 |
| (7) | Research on irrigation and flood control | 0.11 |
| (8) | University research | 0.38 |
| | Total | 7.39 |

Note. Data are for Pakistan, 1966–67. The data are shown in crores of rupees
(1 crore rupees = 10^7 rupees).

This set of vectors of possible allocations, with all vectors having the same sum of components, is in the present example the set D of possible decisions. Without serious distortion, we may think of the set D as containing all vectors satisfying the constraints that every component is non-negative and the sum of the components is 7.39 crores, and thus the decision maker can conceive of continuous variation in the possible decisions open to him.

It is much more difficult to specify even schematically in this example the set S of possible states of nature or the set C of possible consequences, so I shall introduce some drastically simplifying assumptions. To begin with, we shall deal only with the first three components—the allocations to industrial research, atomic energy research, and agricultural research—but it will be apparent that the method of analysis easily generalizes. In restricting ourselves to the first three components of the vector, we

shall assume that the sum $k = 5.66$ of just these three components is constant, and we are considering the allocation of k among these three components.

Keeping these simplifying assumptions in mind, we may now specify the set S of states of nature and the set C of consequences. Let d_x be the decision that allocates vector x . Restricted to our three components, the consequences flowing from any state of nature s and decision d_x are determined, we shall assume, by the equation

$$(1) \quad d_x(s) = (\alpha \log x_1 + \varepsilon_1(s), \beta \log x_2 + \varepsilon_2(s), \gamma \log x_3 + \varepsilon_3(s)).$$

Further, it is assumed that the terms $\varepsilon_i(s)$ are so distributed with respect to the probability distribution of S that $E(\varepsilon_i(s)) = 0$, that is, that their expectation is zero—additional useful assumptions about the distributions of the $\varepsilon_i(s)$ will not be considered here. The logarithmic function for each component, which gives decreasing returns to scale, represents an idealized but not wholly unrealistic assumption. *Post facto* the rate constants α , β , and γ , with $\alpha + \beta + \gamma = 1$, can be estimated by various measures of the output of each scientific sector, ranging from the number of papers published to numerical estimates of the contribution to increased production in the particular sector.

Using equation (1) and therefore implicitly the indicated measure of the utility of possible consequences, we can find at once the expectation of each possible decision d_x :

$$(2) \quad E(d_x) = \alpha \log x_1 + \beta \log x_2 + \gamma \log x_3,$$

with the constraint that

$$(3) \quad x_1 + x_2 + x_3 = k.$$

By familiar methods we can then find the decision d_{x^*} that maximizes expected utility, in particular,

$$x_1^* = \alpha k, \quad x_2^* = \beta k, \quad x_3^* = \gamma k.$$

If we assume that the data of Table 1 represent an optimal decision, we can, as a backward inference, then estimate for 1966–67 the values of α , β , and γ :

$$\alpha^* = 0.34, \quad \beta^* = 0.34, \quad \gamma^* = 0.32.$$

Independent estimates of α , β , and γ along the lines suggested earlier would be interesting even for the present simple model.

3. A second example: United States research allocations for 1960–1970

Data similar to the Pakistani data given in Table 1 are shown in Table 2 for the American allocation of research funds for 1960, 1965, and 1970.

Table 2
United States Federal Obligations for Research, 1960–1970

| Item | 1960 | 1965 | 1970 |
|------------------------|------|-------|-------|
| Life sciences | 511 | 1,167 | 1,533 |
| Psychological sciences | 38 | 103 | 114 |
| Physical sciences | 608 | 1,029 | 1,012 |
| Environmental sciences | | 676 | 575 |
| Mathematical sciences | 25 | 105 | 102 |
| Engineering sciences | 600 | 1,576 | 1,980 |
| Social sciences | 35 | 127 | 200 |
| Other sciences | 33 | 70 | 72 |

Note. Data taken from *Statistical Abstracts of the United States, 1974.*

Exactly the same model can be applied, and similar rate parameters can be estimated on the assumption that the allocation is approximately optimal. Note that the model used in Example 1, when applied here, has the same strong assumption that the utility function is logarithmic in the dollars spent. There are conceptual arguments for using such a logarithmic function as a natural choice to obtain the expected decrease in marginal utility with increasing allocation of funds, but it can certainly be challenged as being too simple an a priori choice to represent the utility of scientific research for any society.

4. Overall allocation to scientific research

Much more critical than the exact choice of a utility function at this stage of discussion are, first, the conceptual basis for evaluating the allocation of research funds to various parts of science, conditional on the total sum allocated for research, and, second, the basis for deciding what percentage of the national budget or the gross national product should rationally be allocated to scientific research.

Let me now turn to the second issue and then return later to the first, of internal allocation.

Derek de Solla Price has made a number of studies of the empirical practices of both developing and developed nations in allocating a proportion of the gross national product to scientific research (see, e.g., PRICE, 1972-73; PRICE and GURSEY, 1975). There is a surprising degree of constancy across nations in this allocation. I shall not review here the empirical evidence.

What I would like to discuss is the tangle of conceptual issues back of trying to make this allocation in a rational way. There are a number of positive arguments for allocating a significant portion of gross national product (GNP) to scientific research. I would like to examine seven of these arguments and, at the same time, emphasize how far each of them is from providing anything like a quantitative guide. After discussing these qualitative positive arguments, I return to the vexing question of how to make the actual quantitative allocation in a rational way.

Argument I. Science for its own sake. There are a number of individuals and a matching number of published arguments saying that science should be supported for its own sake. It is an activity of positive intellectual and aesthetic merit—no further justification is needed. There is something to this argument, but it weighs in at the same level as arguments for studying the history of Etruscan art or establishing a professorship in the baroque music of the late Renaissance. Most of us have positive feelings about Etruscan art and baroque music, and yet not many citizens would feel that a major allocation of national resources should be made to such humanistic studies just for their own sake. Let me be clear on an important distinction here. It is not that there are definitely negative feelings about such studies for their own sake—it is rather that it is recognized that they must be regarded as quite restricted, limited activities whose value is to be thought of in terms of enhancing the broad culture of the citizens. Arguments of this kind would not justify the actual allocations to scientific research characteristic of the modern world.

Let me give a concrete example of an argument that I have had repeatedly. Some of my friends who are mathematicians like very much to make this argument about mathematics for its own sake, but they also find it hard to justify why, in almost every university, the Department of Mathematics is considerably larger than the Department of Music or the Department of Classics. It is characteristic of the temperament of mathematicians that they are not sufficiently intellectually imperialistic simply to say, "But it is obvious that from an aesthetic standpoint mathematics

is simply more beautiful and more engaging than music or the Greek and Latin classics". When pushed, they recognize that departments of mathematics are larger than departments of classics or departments of music for reasons that go beyond the argument of science for its own sake.

Argument II. The importance of understanding as such. A separate and different argument that goes back to Plato and Aristotle concerns the natural human desire to understand the world in which we live. This argument again is at the level of pure science without any concern for application, but it is an argument that separates science from many other studies of the sort just mentioned. For example, it clearly separates science from music, for it is an abuse of concepts to maintain that music leads to an understanding of the world in which we live. Music may lead to pleasure, serenity, or anxiety, but not in any standard cognitive sense to understanding. There is no doubt that this thrust for understanding is a personal motive of considerable importance and value to a great many individual scientists. Again, it is difficult to judge such motives as being anything but positive, but if the realization of such motives were to be the only basis for the justification of allocation of resources to science, almost certainly the allocations would be smaller. Other things being equal, we would all be in favor of scientists' having the leisure and opportunity to pursue one of the highest pleasures known to man—the intellectual contemplation of the universe in which we live. But if none of the rest of us were to benefit from such contemplation, it is doubtful that we would want the allocation of resources to such endeavors to be very substantial. So again the argument is positive but by no means sufficient to guide governmental policy.

Argument III. Source of increased productivity. An argument that is sometimes given before Congressional committees and other places of public testimony is that the results of scientific research are incorporated in technology and lead thereby to increased productivity of the labor force. Such increased productivity leads directly to an increase in the economic well-being of the average citizen. Two of the best classic examples are the introduction of new methods of generating energy, which have become so prominent since about the middle of the 18th century, and the use of scientific research to increase agricultural productivity. A transformation took place that has changed radically the use of manpower from traditional patterns that had been followed for hundreds of

years. The harnessing of various forms of energy is one of the most pervasive and salient features of 20th-century civilization. Although we associate technology with modern electronics in many cases in popular discussions, perhaps the finest example of increase in productivity in the United States is the increase in productivity of agricultural workers. In 1870, the average American agricultural worker was able to produce food for the equivalent of about five persons. A hundred years later, in 1970, an agricultural worker can produce food for the equivalent of about 48 persons, an increase of an order of magnitude in productivity. There are, of course, associated workers in other parts of the agricultural processing industry that need to be accounted for, but, all the same, the increase in productivity is staggering and one of the most significant facts, again, of the 20th century. In the case of both new uses of energy and increases in agricultural productivity, it is fair to say that new scientific understanding was at the heart of the change, even though much of the work was done in settings that did not correspond to modern research laboratories.

A third example concerns the use of computers to transform clerical and administrative work in modern corporate and government bureaucracies. It can be argued that the funds that have been spent for research on computers have already been justified manyfold by the increases in productivity brought about by the introduction of data processing in industry and government. Indeed, it is sometimes claimed, and I think rightly, that the consideration of the transformation of banking alone throughout the world by the use of computers would more than justify the earlier research efforts that led to the creation of modern computer technology. In this case, much of the work was done within the framework of the familiar modern research laboratory.

Argument IV. Source of new products. The impact of scientific research on future productivity can be traced by a continuous analysis of productivity because of the abstractness and generality of the concept. In the case of the argument that basic research leads to the introduction of new products, the argument is more discontinuous and speculative. There is the familiar paradox of predicting new products. If, on the one hand, one were able successfully to predict what new products would be developed by scientific research, then the research would not be needed. On the other hand, it is just in the nature of such research that the new products that are unexpected by-products cannot be explicitly delimited

in advance. The unknown character of the results of future scientific research is one of the best arguments for not attempting to find a precise method for allocating resources. Risks must be taken. If they are not taken, chances for the future will be drastically reduced.

Argument V. Concern for future generations. A related but still different argument concerns the importance of enlarging the base of knowledge for future generations. To take a significant current example, we might argue that really major resources should be put into energy research because future generations will be so much worse off than we are if new cheap sources are not found. To some extent this is already a feature of fusion research: It is, as some of the key current participants have put it, perhaps the first major scientific project whose goal will not be realized during the first generation of scientists working on it.

On the other hand, a wide range of folk, from theoretical economists to shrewd practical investors, will warn against being caught up in an irrational altruistic concern for a potential infinity of future generations. The argument seems wholly persuasive that it is rational to apply a discount rate to concern for the welfare of future generations—it is, of course, another matter to know what the rational discount is, or even if it is possible to have a precise concept. The ethical issues raised by considering the rights of as yet unborn persons have not been at all thoroughly articulated in the literature.

Some positive concern for future generations does seem to be widely shared and constitutes in itself a positive argument for allocating resources to scientific research.

Argument VI. Cost of research. Comparison of the cost of scientific research to other costs, and the comparative cost of particular areas of research, is of importance in the final determination of funds for research. In no country in the world, for example, is the budget for research as large as the budget for education, which is a reflection of the universal demand for education and the priority attached to it in all societies. An example within science is the relative allocation to the physical sciences and to mathematics. As the data of Table 2 show, in 1970 in the United States the ratio was about 10 to 1, but this ratio does not directly reflect the judgment that the physical sciences are an order of magnitude more important than the mathematical sciences, but surely is due in large part to the greater expense of technical experimental work in the physical

sciences as opposed to the mainly paper-and-pencil plus some computer requirements in the mathematical sciences.

Argument VII. Quality of research. Decisions to commit significant funds to research are not infrequently sensitive to the scientific quality of proposals asking for the funding. There is no doubt that the large commitments to space research in various countries have been partly justified by the imaginative appeal of learning more about the universe beyond our planet. At the other end of the physical scale, the same imaginative appeal has buttressed the arguments for large expenditures on high-energy physics.

In comparing various scientific disciplines, opinions about the relative merits of scientific proposals undoubtedly play a significant role, even if not a decisive one, in determining relative allocation of funds. Moreover, in some universities, and even some countries, certain scientific disciplines nearly dominate the research scene and receive what an outsider might regard as a disproportionate share of research funds. As Table 2 shows, in 1960, 1965, and 1970, expenditures for the social sciences in the United States were very much less than those for the biological, engineering, and physical sciences. It would be my conjecture that opinions about the relative merits of past scientific work and current proposals accounted for some of the difference.

5. Potential and limitations of the general model

The central argument for the general model is that it forces an explicit choice of priorities; successful use of the model makes decision making conceptually coherent. Most important, application of the model elicits explicit *partial* beliefs ultimately expressed as numerical subjective probabilities. References to applications have already been given. The general viewpoint of this approach to decision making is prominent in statistics, economics, and certain sophisticated bureaucracies.

There are some serious limitations to the model, however. One is that detailed empirical studies of choice behavior indicate that there are standard biases in probability in individual estimates of the probable occurrence of events. I mention especially the work of Amos Tversky and his collaborators (KAHNEMAN and TVERSKY, 1972; TVERSKY and KAHNEMAN, 1971; also MCGLOTHLIN, 1956). The framework I have outlined does not directly provide concepts for incorporating these results.

Second, in the case of science policy as in the case of almost all setting of social policy, the decisions are made by groups and not simply by single individuals. The framework of decision I have outlined is almost entirely aimed at the theory of individual decision making. Additional complications are introduced by the problems of group decision making, and again additional concepts are needed to deal with the problems of the interaction among the members of a group.

6. Linear models of allocation

A third and more profound difficulty is that we do not have a well-worked-out method to pass from the kind of qualitative considerations just discussed to the general decision model. Because of this absence of anything like an algorithmic technique, it is sometimes argued that we must rely on the intuitive judgment of experienced decision makers. Some difficulties with subjective probability estimates were just mentioned, but there is a more extensive and, for present purposes, a conceptually more important literature in psychology dealing with the comparison of clinical or intuitive judgment and statistical models.

I shall first sketch some of the results reported in this literature and then turn to the question of how these ideas can be applied to our problem of rational allocation. The classical work is Paul MEEHL's *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Literature*, published in 1954. Meehl analyzes a number of studies showing that the prediction of a numerical criterion of psychological interest is almost always done better by a proper linear regression model using numerical predictor variables than by the intuitive judgment of individuals supposedly skilled in such prediction.

A good recent example of this comparison is given in DAWES (1971). The dependent variable was faculty ratings of graduate students at the end of the students' second year; ratings were on a scale of 1 (dropout) to 5 (outstanding). There were 111 students in the sample; the number of faculty members rating each of these students ranged from 1 to 20, with the mean number being 5.67; analysis of variance representing each student as a "treatment" indicated high reliability of the ratings across faculty. The ratings were predicted by a linear regression model using three independent variables: the student's composite score on the Graduate Record Examination, the student's undergraduate Grade Point Average, and a measure of the quality of the student's undergraduate institution.

The multiple correlation of the regression model was 0.38, whereas the correlation of the faculty ratings with the average rating by the members of the admissions committee who initially selected the student was 0.19. Both correlations are low, but the predictions of the statistical model are certainly superior to those of even tutored intuition. Many people find this kind of result surprising in two ways—even though the literature is full of similar findings. First, they are prepared to offer a variety of reasons why the intuitive judgments were not more satisfactory, with the modal suggestion being that the wrong judges were used. Second, they are dismayed at the low correlations, which are typical of demographic analyses. But in fact the causal factors at work may be too numerous and subtle to be analyzed by current methods in a way that will account for more than about a quarter of the variance.

Another objection to the comparison is that the regression model has four free parameters to fit to the data—one for the weighting of each of the three independent variables and one for the additive constant, but the intuitive-judgment model has no free parameters. However, as DAWES (1979) has emphasized, improper linear models, that is, models in which the weightings are not determined by minimizing the sum of a statistical criterion such as squared deviations between the linear combination of predictor variables and the dependent variables, but by fixing the weightings according to some *a priori* scheme, such as equal weights—such improper models do extremely well in comparison with predictions based on intuitive judgment.

What is to be made of these results from a broad conceptual standpoint and what is their implication for decision making? Perhaps the most important conclusion is that human judges are comparatively poor at integrating diverse data, especially multidimensional quantitative data, but much better at understanding what variables are liable to be useful predictors and how their direction of variation will relate to direction of variation in the target dependent variable. The importance of essentially universal agreement on how signs of variation in variables should be related is easy to underestimate. A vast fund of qualitative knowledge is represented in such agreements; it would be difficult to replace intuitive judgments by bootstrapping statistical models that did not use preselected variables.

Problem of application. The pertinent question is how we can usefully apply the ideas just discussed to our problem of rational allocation. In

the first place it should be apparent that it is easier to think systematically about comparative allocations to various major areas of science than it is to compare allocation to social welfare or military defense to that for research.

Let us stipulate the seven factors outlined above—science for its own sake, the importance of understanding, etc. That other factors should be listed I do not doubt, but further analysis in this direction is not important in the present context. Following the work on linear models of statistical prediction, the next step would be to ask a number of knowledgeable persons to rank, let us say as an example for present discussion, each of the major areas of Table 2—biological, psychological, physical, environmental, mathematical, engineering, and social sciences—on a scale of 1 to 7, or 1 to 5, on exactly *one* of the seven factors. Thus each judge is presented with an essentially unidimensional problem—not that the dimensions have a strict mathematical definition. Standard statistical tests of reliability of the measures obtained for a given dimension from different judges can easily be made. Moreover, if the judges are scientists, we could well ask them not to express a judgment on their own area of special interest, in order to eliminate a natural source of bias. Of course, the procedures for selecting expert judges need to be spelled out and are important in any application of the approach I am describing, but I believe such details can be omitted here.

The next problem is to find an appropriate linear combination of those dimensions found to be suitably reliable. One procedure is to ask still another set of judges to weigh the relative importance of each factor and, again testing for reliability, to use the results if good enough. Another approach is to consider the weightings a matter of policy that should be done by appropriate government committees. On the basis of the extensive empirical comparison of statistical and clinical predictions, what is important is to separate the evaluation of each dimension as well as the weightings assigned to each dimension, and then to assemble the results by systematic quantitative methods rather than by attempting one overall intuitive judgment of an essentially multidimensional nature.

Rather than try to sketch the extension of these same methods to the comparison of research fund allocations with other budgetary allocations, I close with some more general remarks about the procedures I have outlined.

I began with the general model of rational decision making under uncertainty, but it is evident that the general model in itself needs to be

substantially supplemented by more specific and concrete ideas to make the model applicable to a complex domain. What I have proposed is to use the results of an extensive psychological literature to structure our problem of allocation in a way that should lead to more rational allocation of resources. The meaning of *rational* in this context is that methods of analysis should be used that lead to more objective and more accurate predictions in those areas of experience that permit comparative tests to be made. This concept of rationality is certainly consistent with that embodied in the general model, but it represents, as I think we should expect and aim for, a more particular sense of rationality relevant to the problem of allocation that has been our focus.

One aspect of the general model that might seem to be lost in the use of linear models for allocation is that of uncertainty, but this loss is only apparent. The appropriate way of thinking of the scaling of each dimension is in terms of expectation, not values, but, of course, *expected* value for the given dimension. This point has not been given enough explicit attention in the literature on linear models of prediction. It would be worth some investigation to determine how feasible it would be to apply some of the experimental methods developed for testing the general model to disentangle the probability and valuation estimates. Moreover, in referring to science for its own sake, sources of increased productivity, etc., it would be interesting to know if experts disagree more on the value of a new development or on its probability of occurrence.

7. Rationality and distributive justice

A variety of literature in economics and philosophy makes the point that issues of allocation must ultimately be judged by criteria of distributive justice. It is worth examining what can be said about distributive justice in the allocation of resources for scientific research. As before, two different questions naturally arise. The first concerns the allocation to individual disciplines, given broad governmental or societal agreement on the total to be allocated, and the second concerns the allocation to research in competition with other areas of demand.

Before considering either of these questions, however, there is one reservation that must be made explicit. We can take as an approach to distributive justice in the allocation of resources the use of market forces to make the allocation. To a large extent this is obviously not the case anywhere in the world in terms of basic scientific research. But when

applied research as well as basic research is considered, then there is a decentralized allocation, partly to be accounted for by market forces, in making the allocation to applied research and occasionally to basic research conducted by private enterprises, especially large corporations, throughout the world. Even within governmental allocation there can be decentralization, because allocations by one government agency can be made in independence of and in ignorance of the allocation made by another agency. I think that the play of such market forces and, even more important, the encouragement of decentralization are factors in the allocations that are actually made that should not be ignored and, in many cases, should be encouraged. All the same, I want to concentrate on appraisal of the total allocation by whatever mechanisms it is made. We can, if we want, think of this appraisal being made from the standpoint of a policy committee not itself concerned with the particular mechanisms of allocation but with judging the appropriateness of the total allocation made by various means, both private and governmental, and by various instrumentalities.

My first thesis is that the internal allocation to various disciplines, once the total allocation to basic scientific research has been fixed, should be according to the rational canons advanced earlier and that considerations of distributive justice add nothing. In other words, the first-order requirement is to use rational methods of allocation, where *rational* has the meaning defined in the previous section. Already here, however, there is a natural reservation in deciding on the criteria to be used to judge the relative allocations, the criteria set forth earlier under such headings as "science for its own sake" and "concern for future generations". Clear issues of distributive justice arise. The criteria should be chosen to satisfy at least some vague criteria of distributive justice. To mention an obvious example, no one could make a case for scientific research if it went against a Pareto principle in some strong way, as, for example, having the results of research clearly be of negative benefit to everyone in society. On the other hand, it is important to stress that the current theories of distributive justice are not rich enough and structurally complex enough to enable us to derive from fundamental principles the more specific criteria for evaluating the worth of scientific research as set forth in Section 4. The gap between general theories and the particular questions of importance in making the allocation is as great as the gap between the general model of decision making and the particular models of linear allocation discussed earlier.

There are some second-order considerations of justice that have not been discussed and yet are of great practical importance in working out actual allocations among disciplines. One is a consideration of continuity. Many would consider it irrational simply to change the allocations in a highly discontinuous way from one year to another because, perhaps, of changes in the judges asked to make the evaluations. Smooth transitions from one allocation to the next, from a practical standpoint, seem essential to satisfy a criterion of justice that is not well articulated in current theories but that goes under the Aristotelian criterion of consistency of judgment.

When we turn to the competitive allocation, as a whole, to research versus other societal demands — for example, for education, for highways, or for welfare — it might seem that criteria of distributive justice could play a more central role, but I am skeptical of any explicit application of current theories of justice to this allocation. As in the matter of internal allocation, criteria of justice can be brought to bear in thinking about and evaluating the detailed considerations set forth earlier, but it seems to me that, except for weak constraints as exemplified in Pareto-type principles, the judgments of the suitability of individual dimensions of evaluation will depend upon the sophisticated intuitions of experienced policymakers. Moreover, I am skeptical that these intuitions can be made theoretically fully explicit and embodied in a systematic theory.

More generally, my skeptical view is that allocations can be rational and can be just, but there is an intuitive gap between general theory and detailed practical decisions that can never be closed by formal theory. This is a fact about human activity that cannot be ignored in our thinking about rationality and justice, whether in general or as applied to particular problems of concern such as those of allocation of resources to scientific research.

References

- DAWES, R. M., 1971, *A case study of graduate admissions: Application of three principles of human decision making*, American Psychologist, vol. 26, pp. 18-188
- DAWES, R. M., 1979, *The robust beauty of improper linear models in decision making*, American Psychologist, vol. 34, pp. 571-582
- KAHNEMAN, D., and A. TVERSKY, 1972, *Subjective probability: A judgment of representativeness*, Cognitive Psychology, vol. 3, pp. 430-454
- KRANTZ, D. H., R. D. LUCE, P. SUPPES, and A. TVERSKY, 1971, *Fundations of measurement*, Vol. 1 (Academic Press, New York)
- LUCE, R. D., and P. SUPPES, 1965, *Preference, utility and subjective probability*, in: Hand-

- book of mathematical psychology, eds. R. D. Luce, R. R. Bush, and E. H. Galanter, vol. 3 (Wiley, New York)
- MARSCHAK, T., T. K. GLENNAN, Jr., and R. SUMMERS, 1967, *Strategy for R & D: Studies in the microeconomics of development* (Springer-Verlag, New York)
- McGLOTHLIN, W. H., 1956, *Stability of choices among uncertain alternatives*, American Journal of Psychology, vol. 69, pp. 604-615
- MEEHL, P. E., 1954, *Clinical versus statistical prediction: A theoretical analysis and review of the literature* (University of Minnesota Press, Minneapolis)
- PRICE, D. DE S., 1972-73, *The relations between science and technology and their implications for policy formation*, Forsvarets Forskningsanstalt (FOA) Research Institute of National Defence (Stockholm. Sweden), vol. 26, pp. 1-31
- PRICE, D. J. DE S., and S. GURSEY, 1975, *Some statistical results for the numbers of authors in the states of the United States and the nations of the world*, in: ISI's Who is publishing in science 1975 annual (Philadelphia)
- RAIFFA, H., and R. SCHAFLER, 1961, *Applied statistical decision theory* (Harvard University, Graduate School of Business, Division of Research, Boston)
- SALAM, A., 1970, *Towards a scientific research and development policy for Pakistan* (National Science Council, Karachi)
- SAVAGE, L. J., 1954, *The foundations of statistics* (Wiley, New York)
- TVERSKY, A., and D. KAHNEMAN, 1971, *Belief in the law of small numbers*, Psychological Bulletin, vol. 76, pp. 105-110

COMMENTS ON PATRICK SUPPES'S 'RATIONAL ALLOCATION OF RESOURCES TO SCIENTIFIC RESEARCH'

BENGT HANSSON

Department of Philosophy, University of Lund, Sweden

1. If I were to summarize what I think is the important lesson to learn from Suppes's talk, I would do it in the following three points:

- (1) There is a gap between formal decision theory and practical societal decisions, especially those on a fairly aggregated level and/or with general political implications.
- (2) Nevertheless, there are important, qualitative observations from the formal theory, which can serve as valuable guidelines for practical decisions, often in the form that they provide warnings for pitfalls.
- (3) One such observation is the inability of the intuitive mind to pay simultaneous attention to all the various aspects of a multi-dimensional problem.

If these are the main points, I cannot but agree without reservations. This, of course, does not preclude that one may have critical view-points on the details of Suppes's argument or on his proposed solution, which basically consists in separating the judgements according to the various dimensions of a problem, and the method for amalgamating these judgements. I will first mention some objections, which I do not hold to be real ones, and after that I will concentrate on three problems which I find relevant. However, these critical remarks should be seen against the background of my basic agreement with what I see as the deeper points of Suppes's paper.

2. It is clear that Suppes's approach is very broad, being concerned only with distributions of gross amounts of money to broad areas of scientific research. It takes no great ingenuity to find that many important problems on the micro level are missing in his treatment.

First of all, there is the problem of a finer division of subject areas. How I rate chemical research from the aspect 'concern for future generations' depends entirely on the kind of chemical research I expect to be done. If it is the continued invention of new types of pesticides for general distribution in the environment, I may give an extremely low rating, but if the research is tuned towards determination of exact biological mechanisms, so as to minimise the amounts needed to achieve a certain effect, I may reverse my judgement and give a high rating.

Secondly, there is the question of which criteria to use when a grant-giving body selects among specific projects. This is not just a variation on a smaller scale of the general allocation problem, because now the skills and abilities of the individual researcher also enter the picture. A closely related question is how the decision-making at lower levels is institutionally organized—centralized or decentralized, split up according to subject matter, geographical units or in other ways, the relative role of active scientists, politicians and research bureaucrats, etc.

These and similar problems are no doubt very important, to a large extent, because they are in some sense close to the individual researcher—they influence his motivation and reinforcement and they define criteria for professional success. Certainly they should not be left out in any comprehensive discussion of the system for resource allocation to scientific research. But the need for a discussion of problems on the micro level must not be admitted as criticism of something which aims only at the macro level. I will therefore rule out this sort of questions as irrelevant to the present discussion.

3. I am afraid that my first real objection is a bit vague. I feel that the idea of marginal utility of research money, or marginal efficiency of research, has received too little attention in Suppes's talk. I have a feeling that there are vast differences in marginal efficiency between different sciences, yet I am not convinced that this fact will be reflected in the procedure suggested by Suppes.

The use of marginality considerations in non-quantitative contexts, especially those with general social implications, is of course subject to controversy, and much has been said and could be said for and against it. I shall make no attempt to discuss all aspects of this problem, but will limit myself to a few scattered observations.

The first one is a rather vague feeling of a difference between the concepts of a *just* and a *rational* allocation of money. To let people rate various

research areas, as Suppes suggests, is to invite an opinion on the merits of these areas—in practice maybe on past importance, but ideally on prospects for future usefulness. But to allocate money in proportion to such ratings, no matter how well-informed and accurate they are, has a flair of distributing rewards in order to satisfy some principle of *justice*—to each according to what it deserves.

A *rational* distribution, however, does not care about deserts or merits, but only about how one should make optimal use of a limited set of resources. It is not a matter of being proportional to something, but of making sure that the money is put where it is most efficiently used. Therefore, rationality *has* to make use of marginal considerations. However, the idea of allocating money in proportion to ratings by knowledgeable persons does not seem to be consistent with such rationality unless at least two implicit assumptions hold: that the marginal usefulness of research money is about the same, or of a comparable order of magnitude, in the various areas; and that it is also fairly constant within one area for various amounts of money.

Both of these assumptions may well be challenged. The first one is perhaps a matter of delicate research politics and I will refrain from discussing it, but the second one can be tackled by conventional economic theory. An important aspect in this connection is the relationship between investments and variable costs.

Investments are usually thought of as pieces of machinery and the like. Indeed, e.g. particle accelerators in physics constitute such investments which are large enough to make a difference even to a national research budget. But there are also other sorts of investments: new research institutes, new departments at existing universities, etc. Such sizeable investments may well create a conflict between rationality and justice: given the amount of money already invested in a certain area, rationality considerations, based on considerations of marginal efficiency, will allocate still more money to that area to make full use of the investments, thereby creating an overall allocation which may well have a definite bias in justice.

Perhaps it would be illuminating to see the difference between e.g. mathematics and physics in this perspective. Once we have got the equipment, it costs rather little to do some more physical research, whereas the cost for mathematical research is almost in its entirety variable.

Apart from the very interesting philosophical problem of whether one can be both just and rational, I think that this is also a very real problem

in the more applied areas of research, such as research on new energy sources. Too large investments are felt to restrict unduly future freedom of action.

Another type of investment, which is particularly important because it cuts across subject boundaries, is that of methodological studies, such as statistics, philosophy of science and certain branches of mathematics. In addition to regarding these as subjects of their own, they may well be treated as investments for other sciences. And they would probably be very profitable investments too, e.g. for the social sciences, but only in a rather long time perspective.

4. One essential point in Suppes's paper is that the intuitive mind is not so good at paying simultaneous attention to all the aspects of many-dimensional problems. He therefore proposes to reduce the use of intuitive judgements to one-dimensional situations, viz. to the ranking of major research areas according to *one* specific criterion.

While I agree that this is a definit improvement, I still think that such a ranking is essentially a *two-dimensional* process, at least for most of the criteria proposed. A ranking of research areas according to e.g. their contribution to increased productivity consists *both* of a judgement of how important the problems of that area are for productivity, *and* one's confidence that the scientists will actually be able to solve the problems. Is it better to tackle a basic and important problem with only a small chance of success, or to reach a small goal with certainty?

This is a paradigm case of a two-dimensional problem. And practically speaking, I think it is a very important one, because it is closely linked to the distinction between natural and social sciences.

Natural scientists typically address well-defined technical questions with a well developed methodology. Even if the individual problems are of limited direct applicability, their solution is reached with a fairly high degree of probability, and the pieces steadily accumulate into something lasting and useful.

Social scientists typically address broad and general questions with a methodology, which only permits uncertain answers, often of limited relevance. They typically 'add new viewpoints' to a problem, but these view-points seldom, or in any case only slowly, build up into a firm, useful and lasting body of knowledge. The problems are not easily split up into smaller pieces with independent possibilities of verification. And while everyone agrees, I think that the problems posed are of the outmost

importance, it nevertheless seems unlikely that we will ever have definite answers.

Maybe this remaining two-dimensionality could be remedied by the same trick once again: let one set of judges assess the importance of the problems and another the probability of success. But I would be sceptical to such a proposal. Since the two dimensions are so closely linked to different areas, the determination of the relative weight of these two aspects would be tantamount to a direct ranking of the natural versus the social sciences, thus being a pre-judgement of the main problem, rather than an analytical clarification.

5. What is perhaps my most important point is that it is not really allocation of funds which is interesting. Promotion of research is the end, and money is only a means. Money, in these connections, usually means some new equipment, salaries and expenses for a few budget years ahead at most, and other short term commitments. But such money is only one of several factors which affect scientific research. The 'production factors' of scientific research could of course be specified in many ways. The following is one which I find useful in many cases:

- (a) trained people, intellectual tradition,
- (b) ideas,
- (c) institutional arrangements,
- (d) money, short term.

All four factors are necessary, and the first step is therefore to make the correct diagnosis as to what factor is the most wanting one. It is not at all certain that it is money, in each individual case.

It is true, of course, that money can help in creating trained people and institutional arrangements, but only in a rather long time perspective. Successful research management therefore amounts to much more than allocation of money — it requires that one solves a rather intricate control problem with long term feed-back mechanisms, and that one constantly monitors the input variables.

Especially in today's society with rapid changes in societal structure and a tendency to increased political control of research priorities, it is frequently the case that new fields of research, or new combinations of research areas, emerge with strong (and justified) demands for society's resources, but where the narrow sector is (a), (b) or (c), rather than (d).

Of course, analysis in terms of narrow sectors can be seen as a refinement

of a marginal efficiency argument. If (d), i.e. short term money, is the narrow sector, we can reasonably expect the marginal efficiency of research money to be comparatively high, but otherwise it will not. Energy and cancer research are two fields, where concern is often expressed about marginal efficiency. Maybe one could dare to venture that the two cases are not entirely parallel, viz. that the narrow sector is (a) for energy research and (b) for cancer research, which, if right, would imply that quite different measures ought to be taken in the two cases.

So the problem is split up into two, like in so many other economic contexts: one is to allocate money for 'production', i.e. for research the next year, the other to determine the level of investment for future research. And the decision on the former problem is to a large extent depending on previous decisions on the latter. This puts restrictions on the problem, which can make it difficult to see proportionality as an optimal solution.

ALLOCATION OF THE RESOURCES TO SCIENTIFIC RESEARCH A COMMENT TO THE REPORT BY P. SUPPES

BORIS YUDIN

*The Institute for History of Science and Technology of the USSR Academy of Sciences,
Moscow, U.S.S.R.*

First, I would like to note the significance of the fact that, for the first time, an international congress of logic, methodology and philosophy of science has embarked on a discussion of the ethical aspects of science. This development testifies, among others, to the growing impact and complexity of the ethical problems arising in modern science as it evolves and to the increasingly stronger links between science and other social institutions, between the life of science proper and the life of man and society.

Besides, the phenomenon is a remarkable one in that it reflects the need for an exactly *methodological* approach to the ethical problems of science. What is at issue is not an approach aiming at no less than settling these problems once and for all and producing some universally binding, uniform recommendations applicable to any situation. In other words, what we have here is not a normative approach. By referring to a methodological approach to the problems of science ethics, I mean, first and foremost, an analysis of the existing, real scientific activity and its social context, an analysis whose subject-matter is the *raison-d'être* or justification of the activity and its prerequisites. It is through such an approach that one can study the mechanisms bringing about the ethical problems of scientific activity and conditioning their nature and specific features. To put it differently, a methodological approach treats the ethical "problematic" itself as inherently linked with scientific activity since the latter represents a type of *human* activity; the approach recognizes the justification, legitimacy and the clearly open nature of this problematic, as well as its historical evolution stimulated both by changes in science itself and in the forms of its societal manifestations.

A methodological approach does not attempt to solve *for* a scientist those ethical problems which he faces in specific situations. In the same way, it does not resolve for the scientist the specific problems of cognition. Instead, the methodological approach subjects these problems to a critical analysis (critical in the sense of Kant) and analyses ways to resolve them, identifying the basis and justification for the decisions taken, which are far from being always realised by the decision-makers themselves.

I think that one of the basic merits of the very lucid and thought-provoking report by Professor P. Suppes is precisely such a *methodological* approach to the problem of resource distribution as applied to scientific research. The model, proposed in the report, is designed to simulate, describe and, eventually, to offer a rational for the processes of taking decisions both with regard to allocation of resources between research institutions and other social needs and with regard to allocation of resources between the various fields of research activities. All this amounts to saying that the problem is to study the possibilities for optimization — albeit partial — of the existing procedures, rather than replacing them with new and, *a priori* fully rational, procedures.

In fact, the making of decisions on these issues, as is often the case in many other similar situations, largely relies on intuitive arguments whose underlying causes are frequently not fully realized. That is why here we have a rather limited scope for formalization or algorithmization, and here one is bound to agree with Professor Suppes who, in his report, takes an attitude of scepticism *vis-a-vis* the possibility of bridging the gap between a general model and detailed practical solutions. Essentially, each such decision blends rational and intuitive components — the question is only that of the magnitude of the rational component and its "terms of reference".

A manifestation of a clearly methodological approach to the problem of resource allocation in Professor Suppes' report is seen in his stressing the need for a continuous, gradual resource allocation, for smooth transitions from a given year's allocation to that of the next. Here, one also has in mind an optimization of the existing practice, rather than its ambitious restructuring, which is apt to bring forth a host of new difficulties rather than resolve the existing ones. (One can note, incidentally, that the report's data on federal U.S. spending in 1960, 1965 and 1970 are far from presenting a smooth pattern if one compares the rates of spending in different research sectors. Here, it would appear that the problem is not so much the lack of justice or rationality as some other factors ac-

counting for the changes in the distribution of spending. It is quite probable, for instance, that the nearly sixfold increase in the spending on social sciences in 1970 over 1960 has been partly due to the student unrest of the end of the sixties.)

I am inclined to see Professor Suppes' central argument in the following. Numerous experimental data suggest that man, in his intuitive judgements, shows a sound awareness of what individual variables can have a tangible impact in terms of inducing variations of a particular dependent variable, and of how the variations of a substantive variable effect changes in the dependent variable. However, man's intuitive judgements turn out far less reliable when there is a need to integrate some variegated or multi-dimensional quantitative data. Consequently, if we identify and process statistically the intuitive appraisals by experts and then build—with reliance on systematic quantitative methods—a corresponding linear combination out of these appraisals, we shall thus obtain a more rational distribution. Here, intuitive judgements will be reserved for a sphere where they are more precise and reliable, namely in dealing with individual variables. Integration or combination of variables, when intuitive judgements become unreliable, call for the latter's replacement by more formal procedure.

Such an approach appears highly interesting and promising, because, among others, it aims exactly at optimising and rationalizing the existing procedures, the product of many years of practice which accumulates the previous experience of decision-taking in resource allocation for scientific research. At the same time, it is clear that materialization of such an approach presupposes solution of many problems. Some of these have been mentioned in the report and it is not necessary for me to go over that ground again. There still remain, however, a number of problems on which I would like to dwell at more length.

The first is the problem of mutual independence of the basic variables, or criteria, underlying intuitive judgements. It may seem that the problem of mutual independence of criteria is of no consequence. This, however, is not so. In fact, if the condition of the independence of criteria is not fulfilled, then an expert appraisal according to an individual criterion may appear unreliable since the appraisal will actually be an *integrated* appraisal. Now, let us rely on this viewpoint to consider those seven criteria for arguments which are proposed in the report for the purpose of rationalizing solutions with regard to allocation of resources for science compared to other societal needs. Proceeding from the same intuitive premises, one can state that, for instance, criteria I and II, i.e. the criterion

of "Science for its own sake" and that of "The importance of understanding as such" are not independent. Same is true with regard to the criteria III "Science as a source of increased productivity" and IV "Science as a source of new products". It is clear that these criteria do not coincide with each other—but the problem is to decide to what extent judgements by experts, say, on criterion III will be independent of their intuitive predispositions with regard to criterion IV.

Let us further assume that we can, using the technique of correlation analysis, determine which of the criteria in the given set are mutually dependent. In this case, however, it will be necessary for us to change the basic set of criteria, which, in turn, poses a new question: what are, generally speaking, the premises we could rely on to select a set of *independent* basic criteria? (Incidentally, it may be remarked that such questions—namely those pertaining to the method of selecting basic units—arise also with regard to other elements of the model proposed by Professor Suppes in his report, on which I propose to dwell somewhat later.)

We can, however, renounce the need for independent criteria or, at least, assume a laxer attitude to it, agreeing that some of the criteria may be mutually dependent to varying degrees. I think that such an approach is quite acceptable. Nevertheless, in doing this we agree to utilize such intuitive appraisals which would refer to integrated, combined basic variables. It would appear, however, that intuitive judgements cannot, in virtue of their very nature, be considered as atomic and mutually independent since they have to do with certain intuitively registered *links* between events, phenomena and processes. In other words, an intuitive appraisal will always remain to some extent integrated, taking some indirect account of both individual parameters and their reciprocal links. And in this respect, namely as regards the choice of basic criteria, the possibilities for a rational analysis within the framework of the model proposed in the report of the problem of resource allocation with regard to scientific research appear to me somewhat more limited than they seem to appear to Professor Suppes.

Another question arises with regard to that treatment of rationality which is contained in the report. By rationality Professor Suppes means utilization of such analysis techniques which lead to more objective and more accurate predictions. On the whole, such an approach to rationality causes no objections. In effect if what is at issue is, say, horse racing, which is one of the examples cited in the report, here one can compare those predictions which assumed the form of bets with the results of

horse racing. In the same vein we can, for instance, appraise the accuracy of predictions if what is in question is weather forecasts.

However, resource allocation evidently belongs to a different class of situations. In this case we cannot compare some subjective appraisals with an objective process which is independent of these appraisals since we simply lack the basis for such a comparison. Resource allocation, when it has been performed, itself defines that reality to which it pertains. That is why it is not clear whether it is possible to consider one of the proposed distributions as more objective and accurate, i.e. more rational in the sense of Professor Suppes, than another. The very notions of objectivity and precision do not, in this case, have a clear bench-mark or reference point. Thus, here we can only speak about rationalizing the *procedure* of decision making rather than rationalizing the *final results*, i.e. the resource allocation proper. Meanwhile there may be no direct dependence of the latter from the former just as an improvement of one of the elements of a system is far from always being conducive to a better functioning of the system as a whole.

In this connection, the question arises as to how substantial the effect of the proposed procedure rationalization can be if it is compared to the final results. It is evident that this question—which is actually a question about the possibilities and limitations of the model proposed by Professor Suppes—touches upon a significantly bigger quantity of factors and dependencies than those which are reflected in the prospect of decision making. At the same time, however, we have so far considered these factors and dependencies only at their qualitative level, in the most general features, without the precision which is made possible at least by the same theory of decision making. In such a qualitative analysis, however, we can begin not with a doctrine of method or a set of techniques existing in a particular sphere, in an attempt to apply them to a given problem, but rather with the problem itself and its essence.

We could point out the fact that the decision-maker taking decisions on allocation of resources for research efforts—be that an individual or an organization—has some notions as to what science is, what its social essence and societal values are, how it is interlinked with social institutions, what societally important objectives can be achieved, and to what extent, by the use of scientific knowledge and methods, etc. It is these notions that largely determine the specific mode of allocation of resources *between* science and other societal institutions and between the various fields of science. Properly speaking, the very question as to what part of society's

resources must be allocated for the development of science as compared to spending on other social needs leads to an analysis of this set of problems.

We shall, further, discover that the most different social groups are interested in how the resources are allocated and that any specific distribution is a function of the interests of each particular group and of the degree of its influence in society. All these and many other factors, taken together, determine the context of decision making, channel or circumscribe it, without, of course, in any way predetermining it unambiguously. It would appear that one of the tasks of a methodological analysis of the problem of resource allocation lies precisely in the need to identify the context, the notions and the premises which influence, directly or indirectly, the allocation of resources for scientific research. This gives us a possibility to make the process itself more rational in the sense that it will, so to say, become more transparent, more accessible to reflexive control, since its aims and the available means to achieve them will become more readily identifiable. In this case, rationality is probably understood in a broader and more classical sense, though, probably, in a less operational sense—for instance, as formal rationality interpreted by M. Weber.

It would appear that such an approach to resource allocation for research which relies on and takes account of certain socially articulated aims is gaining increasingly broader currency in today's practice. The approach materializes through a multitude of the most variegated mechanisms, which, in turn, are themselves being continuously modified and refined. One of the relatively more known mechanisms of this kind is the practice of planning and implementing purpose-oriented programmes. One can note in this connection that in the 1976–1980 period the Soviet Union is implementing over 200 purpose-oriented programmes; future plans aim at transforming purpose-oriented programmes into the chief mechanism of the socio-economic development and, thus, of resource distribution. Each such programme is designed to tackle a specific, societally important task. The tasks may be quite varied, covering such fields as energy, ecology, transport, education, population, etc.; they can also be multipurpose, bearing upon quite different aspects of societal life simultaneously. The process of elaborating a programme helps specify the task pursued—to determine the criteria the expected result is expected to meet, the ways and means to achieve the aim set and the sources and

resources to get them. In the process, there also takes shape a societal organization to implement the programme.

Of importance to us in the present context is the fact that the elaboration of any sufficiently large and important programme provides for expenditures needed to resolve different *research* tasks; in other words, a certain part of the resources within the framework of a particular programme is spent on scientific research. Taking a somewhat different view of the situation, we can say that in today's conditions the spending on science and scientific research represents *something more* than another item of spending similar to allocations for social security, education, development of production, defence, environmental protection, etc.; as a rule, research spending is *also an integral part* of all of these items of expenditure, even if only because the task set is that of a *rational* utilization of the resources channelled into these areas, to say nothing about the fact that scientific knowledge provides the basis for devising means to meet these public needs.

All this poses the need for the society to allocate some of its resources for research; it will become clear that this circumstance cannot fail to be taken into account in an analysis of the way of allocating resources and of the possibilities of rationalizing the process. And here, once again, we place ourselves *vis-a-vis* the problem of identifying some basic premises already referred to in another connection. And in fact, if one is to agree to what was said about resource distribution with regard to purpose-oriented programmes, it may well be that the question of resources spent on scientific research as measured against the spending on other societal needs will prove of quite limited impact. It may be of interest, say, for a statistical analysis to identify the dynamics of annual spending on scientific research or to compare the share of research spending in overall spending in different countries. And still, the amount of this spending will be determined by adding up research expenditures of different departments within the framework of the various purpose-oriented programmes. Data of this kind will, probably, be rather remotely related to the real processes of decision-making with regard to the allocation of resources for scientific research.

All this amounts to saying that the traditional approach viewing research spending as a whole as just another item of spending cannot be taken uncritically as something self-evident. The fact is that before we begin to operate this type of data, it is necessary to find out to what

extent they reflect the current practice of research financing in different countries, different financing systems and the trends of the process.

Here, it would appear necessary to answer the following objection. When speaking about research spending within the framework of purpose-oriented programmes, what one actually has in mind are the resources spent on applied research rather than that part which is consumed by fundamental research. The latter, being an independent area, is not covered by resources allocated for other societal needs. With regard to this argument, one can say the following.

First, I think that allocation of resources for fundamental research has its own specific nature compared to resource allocation for applied research, among others from the viewpoint of the underlying motives and sources of financing. It is not quite clear to me what Professor Suppes meant in his report in the first place—fundamental research, applied research or scientific research in general. One can, however, assume that, of the arguments considered by him, some—like “Science for its own sake”, “Science for understanding as such”, “Concern for future generations”, “Quality of research”—have more to do with fundamental research, while others have more to do with applied.

At the same time, it is clear that the differences between fundamental and applied research cannot be considered absolute. It is known, for instance, that purpose-oriented programmes are utilized for tackling both social problems, i.e. those that are external in relation to scientific research, and fundamental scientific problems, especially if these problems are of an interdisciplinary nature and so multidimensional and complex as not being capable of being resolved by a single scientific research institution.

It was exactly such a purpose-oriented programme that was, for instance, elaborated in the Soviet Union to organize research into reverse transcription-synthesis of the DNA molecules on RNA molecules. Implementation of this programme, baptized “Project Revertaza”, brought together molecular biologists, geneticists, biochemists, physicists and representatives of other research branches from many institutes of the Soviet Union and a number of foreign countries. Due to the interdisciplinary nature of the effort, the programme was financed by different departments: besides the USSR Academy of Sciences, which in the Soviet Union finances a major part of fundamental research efforts, resources also proceeded from other sources, including those allocated for public

health. In the meantime, for the public health system, scientific research spending has above all an applied character.

Thus it will be seen that in this particular case the resources spent to implement the programme cannot be regarded as research spending only since this item of spending is in one list of items of spending on other social needs.

At the same time, it is necessary to point out that all of the above-stated is far from being an advocacy of eliminating the item as useless, or an appeal to do away with research financing independent of other societal needs. What we are trying to point out is that such an independent item may only cover the resources accruing to institutions for whom implementation of scientific research assignments is their chief function. However, the actual share of resources for scientific research is calculated and allocated in a much more complex manner. Roughly, the share of resources which passes through specialized departments exercising control over scientific research financing may be regarded as a source and an indicator of science's autonomy as a social institution, while that share of resources which is received by science from other departments can be represented as a characteristic of the social functions and social inter-relationships of science.

Interdisciplinary problem-oriented scientific research efforts, organized, implemented and financed as programmes or projects, pose, in a study of resource distribution within science, the same problems as those involved in the purpose-oriented programmes—in the study of resource allocation between science and other societal needs. At issue, again, is the adequacy of the basic units. The fact is that if the resources are allocated on the basis of programmes, the question of determining the share of the resources within the framework of a given programme to be used, for instance, by physicists, chemical experts, biologists or by engineering sciences, becomes of secondary importance. Let us assume that such a break-up can be obtained within the framework of each programme or project, and, by adding up data for individual specialities, we shall be able to arrive at an overall distribution of resources between specialities or disciplines. These data, however, will still have only an indirect bearing on the actual decision-making processes in resource allocation inside the scientific research area.

One can, however, come up with a different approach. Let the resources allocated, for instance, for the physical sciences be determined as a sum-

total of resources allocated for research activities of all the physics laboratories, departments and institutes and let us assume that resources are allocated in a similar manner to other sectors of scientific research. Such an approach would be closer to the real practice of decision-making concerning resource distribution. But here, however, we also face certain difficulties.

It is known, for instance, that at present physics and chemistry institutes are studying increasingly frequently problems of a biological nature—such as, for instance, the problems of biophysics and biochemistry, including applied subjects having to do, say, with oncogenic viruses. It is also known, for instance, that biological institutes study genetical and physiological problems having a direct bearing on the human psyche, man's behavior, etc. If, further, we take the category of "environmental sciences", we shall discover that there is, probably, not a single discipline unconnected with environmental problems.

All these factors do blur and complicate the picture of intra-science resource allocation, limiting the reliability of a clearcut and uniform set of categories similar to the one cited by Professor Suppes by way of illustration. It can hardly claim the status of something perceived as unconditional and self-evident.

Let us point out that more or less similar problems are now the subject of intensive study in the fields of science of science and sociology of science. The authors of these research efforts are attempting to identify and separate from each other such units for an analysis of science as a problem area, a discipline, a speciality, etc. It is true that other problems are studied in the process, such as the structure of scientific communications, dynamics of science, the growth of scientific knowledge, etc. However, the data obtained in the process of such a study of the organizational structure of science would, in my opinion, also be of use in the context of a study of resource allocation within science.

The existence of a broad range of interdisciplinary links has another interesting aspect to it. The fact is that, presumably, these links underlie the functioning of mechanisms which exercise a correcting influence on resource distribution within science, thus complementing the traditional mechanisms of resource allocation. Thus, if a physical laboratory is engaged in a study of the biological processes, it means that the resources, in this case in their rather broad sense—including laboratory equipment, technical personnel, research techniques developed in physics, the skills of scientists—are being redistributed between two branches of knowledge.

Such non-formal mechanisms react more flexibly and directly to the trends in the development of science than do rigid formal structures. At the same time, however, they appear to be much less susceptible to rationalisation, and so even their identification and appraisal of their scope pose serious difficulties.

In addressing myself to these issues, I would like to stress once again that the problem is not to give up the study of how resource distribution is effected between scientific research areas and disciplines or attempts to rationalize the decision-making process with regard to such distribution. What I am trying to suggest is that the very premises underlying the decision making process may themselves be the subject of a special study which can shed some light on those factors, links and dependences which influence resource distribution and are, in turn, affected by the latter.

*

In this presentation, my task, as I see it, is much simpler than the one which faced Professor Suppes in his report. There is no need for me to come up with my own model of resource distribution for scientific research, which could claim superiority over the one proposed by Professor Suppes. Besides, I would venture to say that at present we lack sufficient ground to construct a model sufficiently well defined but at the same time capable of describing all the key stages of the resource allocation process. It is my opinion that both the prospect of distributive justice and the prospect of the theory of decision making are too narrow to permit creation of such a model.

A more topical task at present—the state of the art being as it is—would, in my opinion, be an analysis of those substantive premises and notions which underlie the practice of resource distribution and which must be incorporated in a more complete model. Progressing along this path, we shall probably be able to build such a model in the future. I see the exceptional value of Professor Suppes' report in that it makes these premises and notions the subject of discussion and critical analysis, and I would wish to hope that a discussion around that report will give us a better understanding of the processes of resource allocation for scientific research as well as of the factors and conditions influencing resource allocation. It is exactly these premises and notions that can well be the subject for an ethical consideration and an ethical appraisal.

ON GENETIC ENGINEERING, THE EPISTEMOLOGY OF RISK, AND THE VALUE OF LIFE

STEPHEN P. STICH

The University of Maryland, College Park, MD, U.S.A.

I. Backgrounds and beginnings

The early 1970s witnessed a profoundly important breakthrough in molecular genetics. Workers in a number of different laboratories developed techniques which enabled them to remove bits of genetic material (DNA) from a variety of organisms, and to insert this DNA into bacteria in such a way that it became a part of the genetic material of the bacteria. As the bacterial cells duplicated, the transferred DNA duplicated as well. More recently it has been shown that in some instances this transferred DNA is also biologically active; it synthesizes the same products in its new bacterial host as it did in the animal, plant or bacterial cell from which it was originally isolated. The potential practical applications of this new 'recombinant DNA' methodology in medicine, industry and agriculture are so vast that they may in time rival the industrial revolution in their impact on the human environment. For molecular biologists, however, the most exciting applications of the new technology are in the sphere of pure research. Recombinant DNA techniques provide scientists with a singularly powerful tool for studying the basic mechanisms of genetics, especially in the genetically complex cells of higher organisms.¹

Almost from the beginning research involving recombinant DNA techniques was surrounded by controversy. At first the controversy focused on the safety of the research itself. One of the earliest experiments that proposed to use the new techniques planned to insert DNA derived from

¹ For a clear introductory account of the scientific background needed to assess the recombinant DNA controversy, see JACKSON (1979). For further details on potential applications, see CHAKRABARTY (1979).

the SV40 virus into *E. coli* bacteria. But SV40 is known to transform normal human cells in tissue culture into cells which resemble cancer cells, while certain strains of *E. coli* are natural inhabitants of the human intestines. Might SV40 DNA transplanted into *E. coli* ultimately find its way into the gut of a laboratory worker and there cause cancer? Much more frightening was the prospect that a chimerical strain of *E. coli* containing SV40 DNA might establish itself in nature, causing an epidemic of contagious cancer. Though the scenario seemed unlikely, the researchers who were to do the experiments decided to postpone them until the issue of safety could be more widely discussed. Those discussions were at first confined to a very small circle of leading researchers in the field. But as knowledge of the potential of the new methodology spread, and as the techniques themselves became both more powerful and easier to use, two consequences ensued. First, it became clear that the scenario involving SV40 and contagious cancer was only one among many alarming possibilities that we might have to reckon with if all researchers were allowed to use recombinant DNA techniques as they pleased. Second, more and more people were drawn into the discussion about the proper use of recombinant DNA techniques. A voluntary moratorium on certain sorts of experiments was called for, an international conference was held, and the growing controversy spilled over into meetings of university governing boards, city council chambers, and ultimately into the halls of the Congress of the United States.²

In the course of the international debate that followed, issues were raised that were far removed from the original questions about the safety of the research. Would the research yield knowledge that mankind simply ought not to have? Would the technological applications of the new methodology impose strains upon society which we are unprepared to cope with? And if the answer to either question is yes, should steps be taken to stop the research or to channel it in such a way that undesirable knowledge or unwelcome applications would be delayed? Indeed, is it possible to anticipate and control the direction of scientific research?

In the last two years or so, the debate over recombinant DNA has become distinctly more muted. There are fewer newspaper headlines and television interviews. The experts in the field who were once divided and fractious have now, with a few conspicuous exceptions, reached a consensus that the potential dangers of using recombinant DNA techniques

² An accurate and balanced history of the controversy can be found in ROGERS (1977).

are much smaller than some had earlier feared. This emerging consensus and its implications will be of some importance in the remarks to follow. Whatever the reasons, it seems clear that the controversy is well past its peak. Unless we find ourselves confronted with a biocatastrophe, a molecular-genetics-Three-Mile-Island, it is likely that the debate will continue to fade. Yet the recombinant DNA controversy was in many ways the harbinger of debates to come. The moral and conceptual problems that were suddenly made urgent by the global concern about recombinant DNA have applications far removed from molecular genetics. And these problems have not been resolved. They will continue to plague us as we attempt to come to grips with new and potentially dangerous scientific and technological advances.

In an earlier paper I tried to provide a map of the problem terrain that must be traversed if we are to make reasoned decisions about potentially hazardous research.³ What I want to do in the present paper is to revisit two of the more rugged regions in that terrain in the hope of finding a path. The steps I take are very much in the spirit of exploration, and the paths I probe may well prove to be impassable. My motive for sharing these very tentative steps with you at this Congress traces to the conviction that progress in this area will require the work of many minds and the skills of many disciplines. It is my hope that these remarks may prompt others to try their hands at problems whose difficulty is matched only by their importance.

II. Risk-benefit analysis

For many people, the most natural first response to the recombinant DNA controversy is to view it as an issue which could be resolved by the familiar techniques of risk-benefit analysis. Unfortunately, this cheerful view of the issue is very difficult to sustain. For it is far from clear that the techniques of risk-benefit analysis can be applied to this and similar issues. What is more, even if they can be applied, many thoughtful people question whether they ought to be applied, that is, whether we ought opt for the regulatory policy with the greatest expected utility, as this notion is defined in a risk-benefit analysis. For the arguments which follow, it will prove useful to have before us a simplified step by step sketch of

* STICH (1978). STICH (1979) is a slightly different version of this paper.

what would be involved in running a risk-benefit analysis on the recombinant DNA issue. We can then pinpoint the step at which each of the problems we discuss arises.

Step 1: Enumerating alternative policies

The aim of a risk-benefit analysis is to evaluate policies or courses of action, ultimately assigning to each policy being considered a number which will reflect its desirability *vis-a-vis* all the other policies being considered.

So to begin a risk-benefit analysis we must set down a list of policies that we are considering. In the case of recombinant DNA research, one of the policies that must be evaluated is the laissez-faire policy which would impose on recombinant DNA research no restrictions over and above those normally imposed on all research activities. Another on our list of potential recombinant DNA policies would be a complete ban on all experiments using recombinant DNA techniques. Between these two polls there will also be a range of more moderate policies that we want to consider and evaluate. Let us call the laissez-faire policy 'POLICY₁', the complete ban 'POLICY_n', and the various intermediate possibilities that we are considering 'POLICY₂', 'POLICY₃', ..., 'POLICY_{n-1}'.

Step 2: Partial outcomes and total outcomes

The next step in the analysis is to determine a set of outcomes that might result from the adoption of one or another of the policies on the list. Some of these outcomes will be desirable, while others will not. Thus, in the recombinant DNA example, our list of outcomes would have to include such contingencies as causing a cancer epidemic, the creation and establishment of a strain of pathogenic bacteria which are resistant to the current preferred antibiotic therapy, and even the loosing into the environment of a bacterial strain which metabolizes crude oil and gorges itself on the world's dwindling petroleum resources.⁴ The list would also have to include such possible outcomes as speeding up by a decade the discovery of the cause and cure for cancer, the synthesis of chimerical bacteria capable of inexpensively synthesizing medically useful substances such as insulin and interferon, and the development of strains of wheat

⁴ While this example was used with some frequency at the height of the debate, it is singularly implausible. If the imagined bacteria are to change hydrocarbons into water and CO₂, they will need an abundant supply of oxygen, an element not readily found in the depths of an oil well.

or rice capable of fixing nitrogen as legumes do, and thus not requiring artificial fertilizer. Actually, from the standpoint of risk-benefit analysis the possible outcomes I have been enumerating might best be called *partial outcomes*. The point of the label is to stress the fact that a given policy may produce several different partial outcomes. Adopting the laissez-faire policy, for example, might conceivably bring about all of the partial outcomes on our list. Each possible combination of partial outcomes will constitute what we shall call a *total outcome*. So if we have located n possible partial outcomes, there will in general be 2^n possible total outcomes. As an illustration, suppose we restrict our attention to four possible partial outcomes of recombinant DNA policies:

- 1) a cancer epidemic,
- 2) loosing a strain of oil eating bacteria,
- 3) speeding the cure for cancer by 10 years,
- 4) creation of a strain of nitrogen fixing rice.

The sixteen total outcomes that need be considered are then given by Figure 1 (p. 814).

Step 3: Assigning values to total outcomes

Once we have an enumeration of total outcomes, our next task is to assess the value or moral desirability of each of these total outcomes. Sometimes it will be possible to estimate the value of a total outcome by estimating the value of each partial outcome in isolation and simply adding up the appropriate partial outcome values. But often this strategy will not work. The negative value assigned to causing a cancer epidemic, for example, will surely be of a somewhat smaller magnitude if a cure for cancer is also at hand. In assessing the moral desirability of each total outcome, it will not suffice simply to rank order them. For the purposes of a risk-benefit analysis we must assign to each total outcome some numerical measure of its moral desirability which is unique up to a linear transformation. If TO_1, TO_2, \dots, TO_k is our list of total outcomes, then we can denote the numerical measure of the moral desirability of each as $Val(TO_1), Val(TO_2), \dots, Val(TO_k)$.

Step 4: Determining conditional probabilities

In order to use the values assigned to total outcomes in evaluating policies, we must determine what effect each of our proposed policies will have on the likelihood that the various total outcomes will actually occur. So if, as in our simplified illustration in Figure 1, we have 16 possible

| Partial Outcomes | Total outcomes | | | | | | | | | | | | | | | |
|----------------------|----------------|------|------|------|------|------|------|------|------|-------|-------|-------|-------|-------|-------|-------|
| | TO-1 | TO-2 | TO-3 | TO-4 | TO-5 | TO-6 | TO-7 | TO-8 | TO-9 | TO-10 | TO-11 | TO-12 | TO-13 | TO-14 | TO-15 | TO-16 |
| Cancer epidemic | yes | yes | yes | yes | yes | yes | yes | no | no | no | no | no | no | no | no | no |
| Oil eating bug | yes | yes | yes | no | no | no | yes | yes | yes | yes | no | no | no | no | no | no |
| Cancer cure | yes | yes | no | no | yes | yes | no | no | yes | yes | no | no | yes | yes | no | no |
| Nitrogen fixing rice | yes | no | yes | no | yes | no | yes | no | yes | no | yes | no | yes | no | yes | no |

Fig. 1. Constructing total outcomes from partial outcomes

total outcomes, we must determine 16 conditional probabilities for each policy. Let us use ' $\text{Pr}(\text{TO}_i/\text{POLICY}_j)$ ' to denote the conditional probability that total outcome TO_i will occur, given that we adopt POLICY_j . Then the $16n$ probabilities we must determine are indicated in Figure 2.

| For POLICY_1 | For POLICY_2 | ... | For POLICY_n |
|---|---|-----|---|
| $\text{Pr}(\text{TO}_1/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_1/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_1/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_2/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_2/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_2/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_3/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_3/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_3/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_4/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_4/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_4/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_5/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_5/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_5/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_6/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_6/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_6/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_7/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_7/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_7/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_8/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_8/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_8/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_9/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_9/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_9/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_{10}/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_{10}/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_{10}/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_{11}/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_{11}/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_{11}/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_{12}/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_{12}/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_{12}/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_{13}/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_{13}/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_{13}/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_{14}/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_{14}/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_{14}/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_{15}/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_{15}/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_{15}/\text{POLICY}_n)$ |
| $\text{Pr}(\text{TO}_{16}/\text{POLICY}_1)$ | $\text{Pr}(\text{TO}_{16}/\text{POLICY}_2)$ | ... | $\text{Pr}(\text{TO}_{16}/\text{POLICY}_n)$ |

Fig. 2. The conditional probabilities needed to evaluate n policies with 16 possible total outcomes

Step 5: Calculating expected utilities

Once we have assembled the required probabilities and values, we can proceed to calculate the expected utility of each policy. To do this for POLICY_i , for example, we simply multiply

$$\begin{aligned} & \text{Val}(\text{TO}_1) \times \text{Pr}(\text{TO}_1/\text{POLICY}_i), \\ & \text{Val}(\text{TO}_2) \times \text{Pr}(\text{TO}_2/\text{POLICY}_i), \\ & \dots \dots \dots \dots \dots \dots \dots \\ & \text{Val}(\text{TO}_{16}) \times \text{Pr}(\text{TO}_{16}/\text{POLICY}_i), \end{aligned}$$

and then sum the result. A bit more compactly, the expected utility of POLICY_i in our illustration is

$$\sum_{j=1}^{16} \text{Val}(\text{TO}_j) \times \text{Pr}(\text{TO}_j/\text{POLICY}_i).$$

Step 6: Selecting a policy

The final step in a risk benefit analysis is to actually select a policy from among those being evaluated. Conventional wisdom holds that this

is the easiest step: indeed it hardly counts as a step at all. Once we have succeeded in computing the expected utility of each policy under consideration, we need only scan the list and select that policy with the highest expected utility. In the event that two or more policies are tied for first, we can draw lots.

Now perhaps the most striking thing about this sketch of risk-benefit analysis is how enormously arduous a task it would be to actually carry out a full dress risk-benefit analysis in a case where there were a substantial number of policy alternatives and a substantial number of potential partial outcomes. It seems pretty clear that the risk-benefit analysis technique had best be viewed as an idealized strategy for decision making. Often there will be a variety of shortcuts and tricks available which will enable us to make more or less reliable guesses about what a full analysis would select without going through the entire procedure we have outlined. But let us leave to one side questions about the practicality of running a full dress risk-benefit analysis. There are a number of arguments which claim to show that, quite apart from practical problems, risk-benefit cannot or ought not to be applied to issues like recombinant DNA research. It is to two of these that I now turn.

III. Probability, subjective probability and expertise

Experiments and the careful collection of data will clearly be of value in determining some of the probabilities required in step 4 of a risk-benefit analysis. If, for example, one of the policies under consideration would require certain recombinant DNA experiments to be carried out only using specially enfeebled host bacterial cells, then we shall want to know how likely it is that these enfeebled cells might escape from the laboratory after being accidentally ingested by a laboratory worker. To make its escape in this way, the microbe would have to survive passage through the digestive system. And experiments can tell us with some accuracy what the probability of survival is. But there are other important contingencies whose probabilities cannot be easily assessed by experiment. Let me sketch a few examples.

Suppose a pathogenic organism is synthesized in the course of a recombinant DNA experiment, and suppose it does escape from the laboratory. The situation is a potentially nasty one. But just how nasty will depend on whether the organism competes successfully and finds niche for itself in nature. If it does not, then at worst it may infect a handful

of people before dying out. On the other hand, if it does compete successfully in nature, we will have an epidemic, not simply an outbreak. But it is hard to think of an experiment which would tell us the probability that an arbitrary newly synthesized pathogen would compete successfully in nature, and even harder to think of one that we would seriously think of doing.

For a second example, consider the issue of whether a recombinant DNA experiment will result in the synthesis of a new pathogen. Some imaginable experiments may actually aim at producing a new pathogen by, say, removing the gene responsible for synthesizing botulinus toxin from the botulism (*Clostridium botulinum*) bacteria where it naturally occurs and introducing it into *E. coli*. But these are not the cases I have in mind. Rather, I want to consider the possibility that by inserting a normally innocuous gene into a normally innocuous new host we could inadvertently create a new pathogen. What is the probability of this occurring? Here again, there is no experiment that will tell us the answer. Finally, consider the possibility that in the course of a recombinant DNA research program we could damage public health or the environment in some way that we have not yet even thought of. The recent history of DDT and fluorocarbon propellants suggests that this is a contingency which we cannot simply dismiss. Yet there is surely no experimental data we can gather that will point to the required probability. The problem in all of these cases is not simply that a risk-benefit analysis requires knowledge of probabilities that we do not now have. Rather, what is troublesome is that there seems to be no objective, experimental way of determining the probabilities that we need in order to assess the risks and benefits. The apparent impossibility of experimentally determining the required probabilities has led some writers to conclude that risk-benefit analysis is simply not applicable to issues like recombinant DNA. (See, for example, MACKLIN, 1977.)

For two quite different reasons, however, I am inclined to think that this quick dismissal of risk-benefit analysis is a mistake. First, it is often possible to sidestep the problem that some of the needed probabilities are unknown. If, for example, a certain unwelcome outcome would require a sequence of independent events to occur, then the probability of that outcome will be the product of the probabilities of the independent events. If only some of those probabilities are known, we can make a 'worst case' estimate by assuming that the unknown probabilities are very high. We may then be able to show that one policy is preferable to another

on any assumption about the unknown probabilities, including the worst case assumption.⁵

My second reason for resisting the quick dismissal of risk-benefit analysis is one which raises much deeper issues. The problem, recall, is that risk-benefit analysis often requires probabilities which cannot be experimentally determined. But it is far from clear that experimentally determined probabilities are the only ones that ought to be used in a risk-benefit analysis. We can and do often make subjective probability estimates concerning all sorts of contingencies, even though an experimental test of these estimates is, for one reason or another, quite out of the question (cf. SAVAGE, 1954). So why should we not use subjective probability estimates in our risk-benefit analysis?

One answer that has surfaced from time to time in the recombinant DNA debate is that subjective probabilities are guesses, pure and simple, and guesses have no proper place in a serious moral deliberation. This view, which I will call *generalized skepticism* about subjective probabilities, may be construed as making a pair of claims:

- (1) Subjective probability estimates are mere conjectures, 'manufactured numbers' which have no rational foundation.
- (2) Because subjective probability estimates are without rational foundation it would be morally irresponsible to rely on them in deciding how to act. On my view, both tenets of generalized skepticism are mistaken.⁶

To undermine the skeptic's first claim, it may suffice to point out how much of what goes in scientific thinking rests on subjective probability estimates. Consider an example. A paleontologist uncovers the fossilized remains of an animal and, after studying them, draws conclusions about the likely appearance, diet and habitat of the species to which the animal belonged. Some of these may well be couched in explicitly probabilistic terms; our paleontologist may conclude the probability that the animal was omnivorous is about .9. The inference underlying the paleontologist's conclusion rests on a wealth of background knowledge which he himself may be able to articulate only partially. It bears little *prima facie* similarity to an inference about the efficacy of a new vaccine, based on controlled experimental trials. This latter sort of inference is the skeptic's paradigm

⁵ For more on this point, see STICH (1978), pp. 193–196.

⁶ COHEN (1979) appears to be advocating this generalized skeptical position about subjective probabilities. Cf. pp. 315, ff.

of an 'objectively determined' probability.⁷ Nonetheless, it would be simply perverse to suggest that the palentologist's subjective probability is without rational foundation. Inferences like the one in question are prototypical examples of rational inferences, and any explication of the notion of rational inference or rational conclusion must classify the bulk of such prototypical examples as rational. A view which holds most or all of them to be irrational would be most charitably construed as offering an idiosyncratic stipulative definition (cf. GOODMAN, 1951, Ch. 1).

A similar argument can be used to dispatch the skeptic's claim that it is morally objectionable to use subjective probability estimates in deciding how we shall act. Consider the example of a radiologist examining the X-rays of an accident victim. After studying them carefully he concludes that there is probably some foreign object lodged dangerously close to the patient's heart. If he is right, then emergency surgery is called for. But the radiologist says he cannot be sure. His subjective probability that the shadow on the X-rays is caused by a foreign object is only about .8. As in the previous case, the radiologist's probability assessment is not in any obvious sense experimentally determined. Indeed, we know relatively little about the mechanism underlying complex clinical assessments of this sort. On the skeptic's view it would be morally unacceptable to use this subjective probability in deciding how to act. So a surgeon who relies on the radiologist's opinion when he decides to operate would be acting wrongly. Once the consequences of the skeptic's view have been drawn out in this way, however, it is hard to take his position seriously. Surely we all believe that the surgeon has a positive obligation to take account of the radiologist's opinion in deciding how to act, and failing to do so would be a prime example of morally irresponsible behaviour.

My argument so far has been directed against the critic who urges a generalized skepticism about the rationality of subjective probability estimates. But having dispatched that windmill, a more formidable problem looms. I have argued that some subjective probability estimates are paradigmatically rational, and their use in moral deliberations is sometimes morally obligatory. But it is plain that not all subjective probability estimates are rational. People sometimes have foolish subjective probabilities just as they sometimes have foolish beliefs. This fact, lamentable as it is obvious, poses serious problems for risk-benefit analyses on public

⁷ Bayesians would urge that, though the two inferences are *prima facie* dissimilar, they both ultimately rely on subjective probability assessments. Cf. SAVAGE (1954).

policy questions. For the policy making body must decide whose subjective probabilities to utilize in its computation of risks and benefits.

At first blush it might seem that there is a simple and obvious answer to this question. The subjective probabilities that should be used are those of the experts, the people who are best informed and most knowledgeable about the contingency whose probability is being sought. It is my view that this answer is, in spirit at least, the right one. However, it takes only a bit of reflection to see that the prescription to rely on expert opinion is neither simple nor obvious. It gives rise to a pair of questions which are vexing for both the policy maker and the philosopher:

Who are the experts on a given question?

and

Why should we accept their opinion?

There is a clear sense in which the second question presupposes an answer to the first. For the second question asks for a justification of a certain policy, viz. the policy of utilizing the subjective probability of experts in social risk-benefit analyses, and until the first question has been answered, we have not really said just what that policy amounts to. So let us begin by asking what the hallmarks are of an 'expert opinion' worthy of being incorporated into a decision about public policy.

I will start with a bit of history. When the controversy over recombinant DNA research first began to attract a wide audience, in the aftermath of the 1975 Asilomar conference, the molecular biology community was profoundly divided over the probabilities of various gloomy scenarios that had been conjured. Some molecular biologists thought the probability of these contingencies was infinitesimally small, and that concern about them was quixotic. Others thought the probabilities were high enough to merit serious concern and action. Nobel Laureate James Watson, for one, thought the risks so serious that he threatened to seek a court injunction to prevent workers in his labs from doing certain kinds of experiments. (Cf. ROGERS, 1977, p. 43.) Today, however, there is widespread agreement in the scientific community that the earlier fears of some were unwarranted. This consensus is by no means unanimous, of course. Still, many scientists who were initially quite worried about recombinant DNA research have significantly lowered their subjective probabilities of various untoward contingencies. And their new lower probability estimates are shared by many who were not involved with the question earlier on. Watson, ever outspoken, now considers his earlier assessment 'the biggest

damm fool mistake I ever made.' Fully as interesting as the changes that have taken place since 1975 are the events that led to them. Contrary to what might be expected, relatively little new data relevant to the safety issue has been accumulated. Rather, what has happened, for the most part, is that experts have discussed and debated the issue with each other and with experts in related disciplines. In this discussion the participants were reminded of (or informed about) bits and pieces of previously established knowledge from various disciplines which were relevant to assessing the probabilities of various scenarios. Stimulated by the public controversy, the scientific community set about assembling knowledge that bore on the issue at hand. Very little of the knowledge assembled during the debate concerned experimentally determined probabilities of contingencies that would enter into a risk-benefit analysis — the sort of probabilistic knowledge the skeptic demands. But the facts which were brought to the fore did lead many to lower their subjective probabilities.

One of those who did yeoman service in this process of ferreting out, assembling, and circulating relevant information from various disciplines was Professor Davis, with whom I am pleased to share the program today. In a number of articles and lectures Dr. Davis has argued that what we know about the intense competition amongst bacterial strains in nature and what we know about the added metabolic price of replicating a bit of active DNA make it very unlikely that a chimerical *E. coli* strain could establish itself in nature. He has also reminded the scientific community that the common laboratory strain of *E. coli* is itself something of a hot-house plant which is ill suited to survival in the human gut. And he has sought out the data on known cases of laboratory infection and secondary infection. All of this and more has contributed much toward lowering the relevant subjective probabilities of many of those following the debate. (See, for example, DAVIS, 1979.)

At this point, let me insert a brief digression. A number of the observers of the recombinant DNA debate, myself among them, have been critical of the exploitation of the debate by the news media and by people with various ideological axes to grind. However, I am now inclined to think that the controversy and press exposure played an important positive role in the decision making process. It was the notoriety of the issue which provoked experts in other fields, in this case pathologists, epidemiologists, evolutionary biologists and others, to think about what specialized knowledge of theirs might be relevant to the issue, and to contribute that knowledge to the debate.

I am inclined to think that the consensus reached in the community of biologists about the likelihood of various frightening recombinant DNA scenarios provides a paradigmatic example of the sort of expert subjective probabilities that should be utilized in risk-benefit analyses. If this is right, then we can use the case as a model in attempting to specify when expert subjective probabilities should be used. So let me list the features of this case that I take to be important.

1. *Social sanction of expertise.* Who are the experts? The answer, of course, is that different people are experts on different questions. On questions of containment, the experts are biological safety officers and people experienced in handling pathogens in containment laboratories. In assessing the probability that recombinant DNA techniques will speed up our discovery of the causes and cures of cancers, the experts are the leading cancer researchers at the major universities and research institutes or that subset of them who are familiar with the potential of recombinant DNA methodology. On other questions, still other groups count as experts. But why these people? Why not the members of the Society of Automotive Engineers, the mayors of moderate sized cities, or the people who failed out of university molecular biology programs during the last decade? The answer, I think, is that the experts on a given question are the people who make a *socially sanctioned* claim to expertise on that question. This answer is a vague one, of course, since I have not said what this notion of *social sanction* comes to. Though in most cases it will be intuitively clear whose claims to expertise are socially sanctioned and whose are not. Thus, if the question concerns the probability that the pylons on a DC-10 will fail, the socially sanctioned experts are the safety officials of the FAA, not the passengers who offer their views to television interviewers on the evening news. And if the question is the likelihood that laetrile will aid a cancer patient, the socially sanctioned experts are professors of oncology at the leading medical schools, not the people who run laetrile clinics, or health food stores. The socially sanctioned experts on a given issue cannot, in general, be determined simply by polling the members of the society. The names of experts on various questions are generally unknown to the public, nor would most people have much of an idea about the governmental agency, research establishment, academic department, etc., with which appropriate experts will be affiliated. To determine who the socially sanctioned experts are, we would have to work our way up a hierarchy of socially recognized

cognitive authority. So, for example, few people could name the experts on arthritis, or say where they could be found. But if asked how to find out who the experts are, most people would suggest that we ask their family doctor, the head of the local clinic, or perhaps the neighborhood pharmacist. These people, in turn, will not be likely to know who the experts on arthritis are, but they will suggest that we ask a professor at the nearby medical school. These professors will refer us to their colleagues in the appropriate departments. And their colleagues will likely know the names or institutional affiliations of people knowledgeable in the field. But even at this level we have not determined who the socially sanctioned experts are. To do that, we should have to poll the knowledgeable people about whom they consider to be experts. We can summarize the process briefly, if a bit paradoxically, by saying that the experts are the people whom the experts take to be experts.⁸ I think it would be both possible and interesting to say a lot more about the notion of a socially sanctioned expert, but I cannot attempt that project here.

To allay possible misunderstanding, let me stress that I am not claiming that a subjective probability ought to be used in a social risk-benefit analysis merely because it reflects the consensus of the socially sanctioned experts. Several additional conditions must be met as well.

2. Political freedom and the expert selection process. Our understanding of how and why a given group of people come to acquire social sanction for their claim to expertise is very limited indeed. There is, however, a salient feature of the process in the recombinant DNA case that deserves comment. Whatever the social mechanisms may be by which leading professors of microbiology became recognized as experts, they did not involve explicit politically imposed requirements. No group outside the expert community decreed that certain views were unacceptable and others mandatory. The social dynamics by which people became recognized as experts evolved freely without externally imposed political constraints. The situation stands in stark contrast to other cases of socially sanctioned expertise. During the Lysenko period in the Soviet Union, for example, only those who cleaved to the externally imposed party line could hope to gain recognition as an expert in genetics. And in our own time those who would be recognized as experts on various military matters in the

⁸ Note that on this account of what it is to be a socially sanctioned expert, the fact that a man has been awarded a Nobel Prize in medicine and physiology does not entail that he is an expert on any of the issues relevant to the debate over recombinant DNA research.

United States must pass a detailed screening of their current political views and past political activities. I am under no illusions about the difficulties involved in deciding whether (or to what extent) a process leading to social recognition as an expert has evolved freely without external political interference. But these difficulties should not blind us to a distinction which is real and important.

3. Absence of self-interest among the experts. A certain number of the scientists who have participated in the recombinant DNA debate had a deep personal stake in its outcome. Some of these people had plans to do experiments involving recombinant DNA methods and had invested much time and effort in planning the experiments, writing grant proposals, etc. Others had a substantial financial stake in companies established to commercially exploit recombinant DNA technology. If the technology were to be judged too risky, these people would be negatively effected in a clear and specifiable way. However, one of the striking facts about the recombinant DNA case is that many experts involved, on both sides of the debate, had no large personal stake in the outcome. They were people whose own research did not invoke recombinant DNA techniques, and would not do so in the foreseeable future. Nor did they have any financial stake in the outcome of the debate. Some of these people were involved in the debate from early on, and in a number of cases their subjective probabilities about untoward contingencies have significantly declined. (I know of no major figure in the controversy whose assessment of the risks has grown gloomier as the debate went on.) Of course, self-interest comes in degrees and, like political freedom, it is sometimes hard to assess. Still, in many cases it can be (and was) abundantly clear that experts offering their views had little personally to gain or lose.

4. Consensus after an ample dialogue. There are some questions on which most of the socially sanctioned experts share the same view without any sustained discussion, and there are some questions on which the experts fail to approach a rough consensus even after an ample dialogue. But neither of these was the case in the recombinant DNA controversy. There, a rough consensus was reached after a very visible debate in which experts from many different disciplines participated.

Now the thesis I want to urge is this: *If the socially sanctioned experts have reached a rough consensus about the probability of a given contingency, if that rough consensus obtains after an ample dialogue involving experts*

from a variety of disciplines, if the process of expert selection is relatively free of external political control, and if the community of experts includes a significant number of people who are free of conspicuous self interest, then the probability in question ought to be used in social risk-benefit analyses.

Note that I claim to be offering only a sufficient condition for the acceptability of a subjective probability assessment, and not a necessary condition. I suspect there are other subjective probabilities that also ought to be used in risk-benefit analyses, though I am not at all clear just which ones they are. There is a sense in which the view I am advocating is quite a weak one, since the complex condition I have specified applies to relatively few cases, and I say nothing about the case to which it does not apply. Indeed, I once thought my thesis was so weak that it would be utterly uncontroversial. But in this I was clearly wrong. For despite all the hedging, my view is clearly an 'elitist' view at odds with the populism currently fashionable in some circles. When the conditions specified have been met, I claim that we ought to accept the consensus of the experts about probabilities, and if the 'man in the street' has any views on the issue, we should ignore them. Plainly this is a view which requires some defense. So let us now ask why we should, under the conditions specified, accept the subjective probabilities of the experts.

I suppose the nicest defense of the policy of accepting expert opinion would be one which showed that when the conditions I have specified have been fulfilled the experts generally have gotten the right answer. If we could establish this, or even the weaker claim that the experts have been right more often, on the average, than any other identifiable group, then we might try the following argument: Since the experts have been right more often than anyone else in the past, this is reason to think they will continue to be right more often than anyone else in the future. Unfortunately, however, I am unable to make any serious case for the basic premise of this argument. If there are data showing that expert subjective probabilities have been more accurate than those of other groups, I certainly do not know of them. I confess to being rather skeptical that anyone has such data. Indeed, on second thought, it is not at all clear that such data would do much to justify our reliance on expert opinion, or that data to the contrary would in the least undermine this reliance. For consider the following imaginary experiment. Suppose we set out to test the hypothesis that experts have been right more often than any other group. We randomly select a large number of issues where we are very confident that the right answer is now known. We then randomly

select historical periods to pair with each issue, taking care that there were socially sanctioned experts on the issue in the historical period with which it is paired. We then set a team of historians to work comparing the track records of the experts with those of other identifiable groups. I think we all suspect that the experts would win hands down. But suppose that, much to our surprise, they did not. Suppose that our team of historians turned up a sect of mystical rabbis in Jerusalem which, for the last two millenia, has been taking heterodox stands on all sorts of empirical issues of the day. And while both the rabbis and the experts have made their share of mistakes, the track record of the rabbis is distinctly better. Would evidence like this be a good reason for no longer accepting expert opinion, or for relying on the rabbis instead? I think the answer to both questions is clearly no. If this is right, then our justification for relying on experts has little to do with their past success. But if we cannot justify relying on expert opinion by citing their previous successes, why should we bet on the experts?

I am inclined to think that this question is a close cousin of a very venerable philosophical problem, the problem of induction. As I view it, the problem of induction asks why we should accept certain principles of non-deductive inference, viz. principles which after suitable reflection we are inclined to endorse as rational.⁹ When the problem of induction is cast in this way, our problem about expert subjective probability is not merely a relative of the problem of induction, it is a special case of it. For one of the principles of non-deductive inference that I believe many of us are inclined to endorse as rational after suitable reflection is the principle of accepting expert subjective probabilities when the additional conditions I have specified are met. If I am right, then it should not be at all surprising that the past success of experts does not justify our reliance on them, for notoriously, past success is of no use in justifying more familiar inductive principles. The imagined historical ‘experiment’ of the last paragraph also has an illuminating parallel in the case of more standard inductive rules. Suppose we set out to test the historical track record of the canons of induction we take to be rational, and discovered that, while their record was pretty good, there is a quite different set of inferential rules which would have done even better. Would such a discovery count as a good reason to abandon our standard inductive rules?

⁹ Cf. GOODMAN (1955), Ch. III. For some reservations and elaborations, see NISBETT and STICH (to appear).

The answer had better be no, since in this case we know *a priori* that, since the standard rules do not always get the right answer, there is bound to be an alternative set with a better track record.

Here the argument over accepting expert subjective probability can take one of two paths, both of them short. First, an opponent may simply deny that after suitable reflection he is inclined to accept this principle as rational. For such an opponent I would have no persuasive reply. He is, I suggest, analogous to the person who uses terms like Goodman's celebrated 'grue' to describe the world and make predictions about it. An object is grue if it has been examined prior to now and found to be green, or if it has not been examined and is blue (cf. GOODMAN, 1955, Ch. III). Presumably all emeralds hitherto examined have been green, and this is the reason we believe the next one will be. But all emeralds hitherto examined have also been grue, and the person we are imagining infers from this evidence that the next emerald will be grue as well. Prior to unearthing the next emerald, there is no rational way of changing his mind. Of course, we are all inclined to think that a person who consistently invoked terms like 'grue' in his inductions would not be a good risk for a life insurance policy. I think much the same is true for the person or the society which rejects expert authority in favor of some more democratic way of determining probabilities.

The second path for the argument to take is for the critic to agree that accepting expert subjective probabilities (under the conditions specified) is rational, then go on to demand some reason why we should do it. His analog in the case of more familiar inductive principles is the person who concedes that a given inductive canon is rational, then demands some further reason why we should use it. The only response, I think, is the Wittgensteinian observation that justifications have come to an end.

Before leaving the topic of subjective probability and expertise, let me forestall a possible misunderstanding. The view I have been urging is that, under certain conditions we ought to accept the consensus of experts about probabilities. But this is not to say that we should accept the consensus of experts on whether a risk is worth taking or even on whether the expected utility of a given policy is greater than the expected utility of another policy. For to answer these questions we must know more than the probabilities that various policies will lead to various outcomes. We must also know how these outcomes are to be valued. And on that question the views of people who are experts in one field or another carry no special weight.

IV. The value of life

In the outline of risk-benefit analysis presented in Section II, the question of how outcomes are to be valued arises in Step 3, and it will surely come as no surprise that this step bristles with difficulties. In the present paper I want to focus on just one of these, though a central one. Once again, the problem can be illustrated in the context of the recombinant DNA dispute. Some of the desirable partial outcomes that may result from a policy of encouraging recombinant DNA research would involve a very substantial economic gain for certain individuals and corporations. So, for example, one of the proposed applications of recombinant DNA techniques is the construction of a bacterial strain that would selectively absorb and retain valuable metals like platinum which are left in very low concentrations in certain industrial waste products. If such a bug could be constructed, it would be economically feasible to recover this platinum. However, the research leading to this pleasing outcome is not totally free of risk. There is some chance of inadvertently creating and releasing into the environment bacteria which will do substantial ecological damage, or even cause illness and death to humans. Now in deciding how we shall rank our outcomes, the total outcome including the platinum concentrating bacteria and, say, a dozen deaths caused by the research, will surely be ranked lower than the total outcome with the platinum eating bug and no deaths. But how much lower? Without an answer, our risk benefit analysis simply cannot be run. But to give an answer is, in effect, to assign a monetary value to human lives. And what could that value possibly be?

Some writers have urged that since risk-benefit analyses on policies involving life and death generally require that we assign some economic value to human lives, it is morally repugnant to use risk-benefit analyses in making these policy decisions. On their view, the value of a human life simply cannot be reckoned in economic terms, and the attempt to do so reflects a callous disregard for the fundamental principle of the sacredness of human life. However, I think that a bit of reflection will convince us that this generalized refusal to assign an economic value to human lives is not a tenable moral stand. It forces us either to make absurd moral decisions, or to assume the posture of a moral ostrich and refuse to consider difficult moral questions at all. To see this, consider the example of automobile safety requirements. It would be easy enough to build vehicles which could withstand head-on crashes at their maximum

speed without causing serious injury to their passengers. I would guess that most military tanks are examples of such vehicles. It is beyond dispute that if we required all the vehicles on our public roads to be built like tanks, thousands of lives would be saved each year. Yet no one seriously urges that we should require passenger cars to be built like tanks. The reason is obvious enough: to do so would be prohibitively expensive. We are simply not willing to spend that much money to save those lives. Now for those critics who find it morally unacceptable to assign any economic value to lives there appear to be only two choices. They can advocate building cars like tanks no matter what the cost, in effect assigning lives an infinite value, or they can wring their hands and refuse to say when a proposed safety regulation would cost too much. Neither alternative has much appeal.

I have been arguing that it is not morally objectionable to ask how human lives are to be weighed against economic considerations. But even if this point is granted, we have made no progress with our original question: How much is human life worth? How are we to find out?

Lest I raise false expectations, let me say now that I have no numerical answer to offer. I will not even try to say what a life is worth in dollars or Swiss Francs or ounces of gold. What I shall offer are a few reflections on why the question seems so hopelessly difficult, and how to make it easier. In brief, what I will urge is that the value of life question appears intractable because it is the conflation of many morally different questions. There is no single economic value to be assigned to a human life, but many different economic values depending on the details of the case in which the question arises. On first hearing, many people are inclined to accept a naive principle of equality which holds that when life and death are at stake, all people are to be counted equally under all circumstances. Thus if dollars must be balanced against lives, all lives must be valued the same. It is my view that this unsubtle egalitarianism will not endure scrutiny. There are many differences among persons and circumstances which morally justify valuing lives differently. To make the case let me list three of the differences that I think are most compelling.

1. *The quality and quantity of life saved.* Suppose we are policy makers considering a pair of additions to our national health service. The first program is aimed at improving prenatal care, and the consensus of the experts is that the program will save the lives of a thousand infants a year, infants who without the program would die in early childhood. The second

program proposes to treat a disease of the aged, and expert opinion holds that the program will save the lives of a thousand pensioners a year who would otherwise succumb. How much would we be willing to spend on each of these programs? The naive egalitarian holds that the answer must be the same in both cases. But most of us, I suspect, are inclined to think that the prenatal program merits a significantly greater expenditure. Analogously, a program which *cures* people of an otherwise fatal disease merits a much greater expenditure than one which keeps them alive but leaves them incapacitated by intense and enduring pain. In each of these cases, it is easy enough to say what the morally relevant difference is between the hypothetical health care programs. In the first it is the quantity of life that has been saved, measured in days or years, and in the second it is the quality of the life that has been preserved. It is, I would urge, morally defensible to take account of both factors in weighing the economic value of a life, and under some circumstances it is morally indefensible to ignore them.

2. *Known victims vs unknown victims.* It is often noted with considerable irony that our society is willing to spend much more money to save the life of a person whose identity is known than to save the life of a person whose identity is unknown. Thus, for example, we consider it appropriate to spend millions of dollars to rescue a handful of coal miners trapped after a mine explosion, but we are reluctant to spend an equal amount of money on mine safety, even though expert opinion assures us that such an expenditure would, over the course of a decade, save substantially more lives than are saved by our rescue operation. For a naive egalitarian, such decisions are morally unacceptable; we should not value one coal miner higher than another simply, because we know his name. On my view, however, there are morally significant differences between these cases, differences which may well justify spending more on the known victim than on the unknown.

It is a psychological fact of considerable importance that a decision which harms or helps a specific person whose identity is known to us is often substantially more traumatic or gratifying than a decision which will harm or help unknown persons. Thus it is no accident that the charities which help children in underdeveloped countries often pair off a donor with a needy child, arranging for the child to correspond with the donor and send him a picture. The personal relationship thus estab-

lished provides strong motivation for the donor to continue his contributions. The donor finds it a much more enriching experience when he knows who is being helped by his contribution and how. On the other side of the coin, it can be a psychologically devastating experience to make a decision which condemns another person to death when we know the identity of the person. And this anguish must be taken account of in reckoning the moral costs of deciding that a life shall not be saved. When the decision is a public one the anguish may be multiplied many times over, since many of those in whose name the decision has been made will feel a sense of responsibility for the act, and will share some of the decision maker's trauma.

It is not only those who feel responsible for a life or death decision who react differently when the identity of the victim is known. The victim's friends and relatives react differently as well. In mine disasters, for example, it would not be unusual for the family of a man who was killed in an explosion to feel considerable hostility toward the mine operators and safety inspectors. However, we would anticipate much more intense and potentially explosive hostility if trapped miners are allowed to die because a rescue operation would be too costly.

Now the defender of the naive egalitarian principle may urge that the psychological differences we have noted are in some way irrational, and that people ought not to react as they do. (Actually, I am very skeptical indeed that any case can be made for this view.) But even if the egalitarian is right, the fact remains that people do react differently when the identity of the victim is known to the decision makers, and there is little hope of persuading them to react otherwise. Thus in our moral deliberations, the death of a victim whose identity was known in advance ought to be counted as a greater calamity than the death of a victim whose identity was not known (other things being equal), because the former death brings with it a greater burden of anguish to the decision makers and hostility to the bereaved.

3. *Voluntary risks and involuntary risks.* People often choose to engage in a risky activity knowing full well that the choice may lead to their own death. Hazardous sports like mountain climbing and parachute jumping are good examples of such activities. However, not all of the risks we are subjected to are undertaken voluntarily. In many parts of the world the very act of breathing the air carries with it the risk of causing cancer,

emphysema and a variety of other diseases. Those who would avoid the risks of mountain climbing can simply stay on level ground. But there is no comparably simple way to avoid the risks of breathing.

Now suppose that a pair of public policies have been proposed, one to make mountain climbing safer, the other to make breathing safer by reducing pollution. Suppose further that the relevant experts agree that each policy can be expected to save about the same number of lives each year. How much would we be willing to spend on implementing each policy? The naive egalitarian holds that whatever the sum, it must be equal in each case. But most people are inclined to think that the pollution control project merits a significantly greater expenditure. More generally, I think we are inclined to discount the cost of a death to the extent that the person freely and knowingly undertook to endure the risk which led to his death. This principle raises some intriguing questions about how the degree of freedom of a decision is to be assessed. Are employment related risks, for example, undertaken freely? Intuitively I am inclined to say that the Hollywood stunt man's decision to accept the risks of his profession is considerably freer than the coal miner's decision to accept the risks of his. Building a theory which will capture such intuitions is a formidable task which I happily leave to others.

I think it would be possible to add quite a number of items to my list. But the three I have already mentioned should suffice to establish the point I am arguing. That point, recall, is that the naive egalitarian is mistaken. We need not assign a single value to all human lives in all risk-benefit analyses. There are morally respectable reasons for assigning different values to different lives, and for assigning different values to the same lives under differing circumstances. Thus the task of assigning an economic value to a human life in a risk benefit analysis is actually many different tasks, all of them difficult. It is not, as some have feared, a single task which is *prima facie* impossible.

Before leaving the topic of assigning economic values to lives, let me make note of a confusion that has, of late, had profoundly unhappy consequences. There are many contexts quite removed from risk-benefit analyses where it is necessary or appropriate to assign an economic value to a human life. One of these situations is in a court of law when one person is found at fault for the death of another, and the question to be settled is how much compensation should be paid to the family of the deceased. In deciding what just compensation would be, many issues of social policy come into play. However, contrary to what a naive egalitarian

might urge, there is no reason to insist that the sum which would be a just compensation for the family would also be the appropriate sum to use in a risk-benefit analysis. It appears that the Ford Motor Company made the naive egalitarian's mistake in calculating whether it was worth-while to redesign the gas tank on the ill fated Pinto. After estimating (with some accuracy) how many deaths would be avoided by moving the gas tank, and how much such a move would add to the cost of building the car, they calculated that the cost per life saved would be greater than the average compensation for a death awarded by the courts in recent years. Thus, they decided not to move the gas tank. Public outrage over Pintos which burst into flames after rear-end collisions ultimately convinced Ford that it had used the wrong number. But many people paid with their lives for Ford's naive egalitarianism.

V. Conclusion

In the pages that precede we have looked at a pair of arguments which, in very different ways, try to establish that public policy concerning risky science and technology cannot or should not be determined by risk-benefit analysis. In both cases I have found the arguments wanting. Neither the difficulties posed by probabilities nor the unwelcome prospect of assigning dollar values to lives is reason enough for abandoning the effort to evaluate risks and benefits. I confess that I would have found the opposite conclusion profoundly depressing, since on my view there is no rational, morally responsible alternative to using risk-benefit analysis in these cases. It hardly needs to be said that I have not had the last word on either of the problems I have considered. And there are other problems, equally vexing, on which I have had no words at all.¹⁰

References

- CHAKRABARTY, A. M., 1979, *Recombinant DNA: Areas of potential applications*, in: JACKSON and STICH, 1979
COHEN, C., 1979, *On the dangers of inquiry and the burden of proof*, in: JACKSON and STICH 1979

¹⁰ I am grateful to Richard Nisbett and David Jackson who read and criticized earlier drafts of this paper. Many of the ideas in Section III were evolved in the course of discussions with Richard Nisbett, and he deserves credit (or blame) for anything in that section that he does not disagree with. My research was supported by the U. S.-U. K. Educational Commission and by the American Council of Learned Societies.

- DAVIS, B. D., 1979, *Evolution, epidemiology, and recombinant DNA*, in: JACKSON and STICH, 1979
- GOODMAN, N., 1951, *The structure of appearance*, 2nd ed. (Bobbs-Merrill, 1966)
- GOODMAN, N., 1955, *Fact, fiction and forecast*, 2nd ed. (Bobbs-Merrill, 1965)
- JACKSON, D. A., 1979, *Principles and applications of recombinant DNA methodology*, in: JACKSON and STICH, 1979
- JACKSON, D. A., and S. P. STICH, 1979, *The recombinant DNA debate* (Prentice-Hall, Inc. Englewood Cliffs, N. J.)
- MACKLIN, R., 1977, *On the ethics of not doing research*, Hastings Center Report, 7, 6 (December)
- NISBETT, R. N., and S. P. STICH, *Justification and the psychology of human reasoning* (to appear)
- ROGERS, M., 1977, *Biohazard* (Knopf, New York)
- SAVAGE, L. J., 1954, *The foundations of statistics* (John Wiley and Sons, New York)
- STICH, S. P., 1978, *The recombinant DNA debate*, Philosophy and Public Affairs, 7, 3
- STICH, S. P., 1979, *The recombinant DNA debate: Some philosophical considerations*, in: JACKSON and STICH, 1979

ALLEGED THREATS FROM GENETICS

BERNARD D. DAVIS

Bacterial Physiology Unit, Harvard Medical School, Boston, MA, U.S.A.

Though the scientific community has long enjoyed virtually complete autonomy in directing and regulating its research, in recent years demands for external regulation have mounted. In the physical sciences these demands arose primarily from belated recognition that large-scale technology has yielded not only benefits but also costs and dangers, ranging from the threat of nuclear catastrophe to despoliation of the environment. Accordingly, attention has been focused on the unquestionably real problem of controlling harmful technological applications. In biology, in contrast, it is largely conceivable dangers in the basic research itself and in its future consequences, rather than demonstrated dangers in its present applications, that have caused most concern. I shall consider three areas of concern, all involving genetics: the possible generation of dangerous products, powers, or ideas. For the first two recombinant DNA serves as a paradigm. Here Professor Stich has already emphasized the conjectural nature of the hazard, and the need for forming judgments based on subjective probability. I agree entirely with his analysis, and I shall focus on the pertinent scientific considerations, and on the political-scientific question of how such matters might be better handled in the future.

Dangerous products

The first concern, over possibly dangerous products of genetic research, exploded with the development of a procedure that made it possible to insert small blocs of DNA from any source into bacteria. A decade ago such a powerful tool for studying living systems would have been greeted solely as a remarkable breakthrough. In the altered current atmosphere,

however, public discussion has focused much more on the conjectural danger of creating novel organisms that might harm the public or the environment.

This possibility was initially raised, in 1974, by a group of molecular geneticists; and while their concern was very much in the tradition of a responsible scientific community, they departed from tradition—perhaps in reaction to years of criticism of scientific elitism—by making their initial concern public before the matter had been explored in private. Though this action, and the further request for governmental regulations, at first received acclaim, a handful of other scientists, for various reasons, were not satisfied with the restrictions and expressed great alarm over such research. This man-bites-dog news inevitably had great appeal for the news media, and widespread public anxiety soon followed.

Meanwhile, apprehension has largely subsided. Let me briefly summarize some of the reasons. (1) After several years of work with such recombinant bacteria, in many dozens of laboratories, the hazards remain entirely conjectural: no illness, and no environmental damage, has been traced to this source. (2) Mutant host strains have been developed that provide a huge increment in safety: they can be grown in the laboratory, but they self-destruct rapidly under conditions encountered in nature. (3) The increasing evidence for promiscuous uptake and transfer of DNA in the bacterial world makes it extremely likely that in nature bacterial cells in the mammalian gut occasionally pick up DNA not only from other bacteria but also from surrounding host cells; hence the new laboratory recombinants would not be a novel class of organisms after all, but would be additional examples of a class that evolution has dealt with all along. (4) Evolutionary principles, which entered the discussion late, predict that spread of any harmful novel microbes would involve much more than their ability to multiply. Thus natural selection of an organism depends on its adaptation to the environment, and adaptation, in turn, depends not on the properties of a single gene but on a balanced set of genes; hence only an infinitesimal fraction of genetic novelty in nature survives. Since genetic balance is hardly likely to be improved by insertion of DNA from a distant source, bacteria containing such insertions will almost certainly be at some disadvantage in competition with naturally adapted organisms. (5) It was also recognized belatedly that the problem is fundamentally one in epidemiology, and the input of persons knowledgeable in this area turned out to be highly reassuring. Thus recombinant *E. coli*, with 0·1% inserted foreign DNA, will retain the limited habitat

of *E. coli* (i.e. the vertebrate gut) and the mode of spread of *E. coli*. It is therefore pertinent that spread of strains of this organism, and of related enteric pathogens such as the typhoid bacillus, is easily controlled by modern sanitation. Moreover, even with kinds of pathogens whose spread is much harder to control, such as respiratory viruses, no large epidemic has ever arisen from a laboratory. In a word, no expert in infectious disease held that any recombinant *E. coli* could be as probable a source of spread of disease as the many well adapted, highly virulent organisms that already exist and that may be encountered in any diagnostic laboratory at any time.

Eventually, then, it was recognized that the discussion had been misguided by being framed in terms of the absolute question: "Can you prove that the following catastrophe could not occur?" As in every other interaction with the real world, a rational response must deal with probabilities. Public anxiety subsided, the NIH Guidelines were relaxed somewhat (and the responsible committee has recently recommended much further relaxation), and the U. S. Congress concluded that the issue was too fluid and complex for rigid legislative regulation. Nevertheless, an expensive bureaucracy has been set up, and it may be with us for a long time. The greatest cost may be the precedent for future further bureaucratic interference with the responsible role of the scientific community in assessing the value and the risks of novel capabilities in research.

What lessons can we learn from the episode? First, it is clearly legitimate for the public to be concerned about the production and use of dangerous materials in research. Where the dangers are well defined, as with inflammable or toxic chemicals, there has been no real tension between the public and the scientific community, or within the latter group. But where the potential dangers are matters of judgment rather than demonstration the value of informed judgment must weigh heavily, as Professor Stich has emphasized. Moreover, there is bound to be dissension in the scientific community during the early phase of digesting the relevant facts and principles. I would question whether the public benefits from participating, or from having the media present, during the highly technical phase. On the other hand, some form of participation by the public (usually through its representatives) is essential in the later phase of risk-benefit analysis, which must follow when risks are judged to be significant.

Second, while we are often faced today with the assertion that no group can be trusted to police itself, I would point out that in certain areas the

scientific community has long been given that trust: the advancement of its goals was believed to coincide, and not to conflict, with the benefit of society, and policing could only hinder that advancement. In the recombinant DNA issue, where this trust was lost, it is now clear, in retrospect, that the scientific community had indeed acted thoroughly within its tradition of social responsibility, and the recent public mistrust resulted in a futile and enormously expensive exercise—expensive not only in time and money, but also in terms of the morale of the scientific community and the public image of science. Accordingly, while there is great appeal in the principle of public participation in any discussions that ultimately involve the public interest, perhaps this principle, like other principles of social action, has its limits.

I would conclude then that there clearly is room for procedural improvements in separating the technical assessment of the hazards of research from the political pressures that attend the subsequent process of decision-making. There also is need for improved communication with the public—but this problem is complicated by the way the media selects news. For I can attest that during the long period of intense publicity over the alleged dangers of recombinant DNA it proved impossible to get even some of the most responsible news organizations to pay attention to the reassuring arguments from epidemiology and evolution.

Dangerous powers

A second area of concern is knowledge that might give us dangerous powers. In the physical sciences such concerns have ranged from military applications of nuclear energy to electronic invasions of privacy. In the biomedical sciences similar concern has arisen over genetic engineering: a phrase that has frightening overtones. However, we must recognize that in medicine genetic engineering means gene therapy: the replacement of the single defective gene that is responsible for a hereditary disease. Over 2000 such diseases are now known (most quite rare), and in over 200 (such as *sickle cell anemia* or *phenylketonuria*) the abnormal gene has been biochemically defined.

Such replacement of a defective gene, considered by itself, is surely as legitimate as the continual replacement of gene products, such as insulin. But 10 years ago a group of political activists working in science asserted that if advances in molecular genetics achieve this goal the same techniques will be used to manipulate personalities. This widely accepted

proposition clearly added to public anxiety in the recombinant DNA debate—though more as an undercurrent than as a central issue.

Such a prospect would indeed be frightening, if it were realistic. However, the relevant technical facts are highly reassuring. First, therapy even of single-gene defects still seems, unfortunately, very far off, except for defects in those cells that function in widely distributed, loosely organized, locations (i.e., the precursor cells in the bone marrow that give rise to blood cells). Second, even if the more difficult goal of gene therapy for defects in localized organs (e.g., the liver) should eventually be achieved, an enormous technical gap would still separate it from any useful, predictable transfer of behavioral genes. For the organ of behavior, the brain, is at the other end of the spectrum from the blood cells: its characteristics depend primarily on the specific connections between its 10 trillion switches. And though we know nothing in molecular terms about the genes that contribute, via differences in brain structure, to individual differences in behavioral potentialities, we can be sure that an enormous number of genes are involved in guiding the development of the circuitry of the individual's brain. Moreover, these genes will have done their work before birth: gene transfer could not conceivable rewire an already developed brain.

For these reasons genetic manipulation of personalities will no doubt long remain remote: meanwhile many other means of manipulation are already at hand. Nevertheless, if attacks on genetic engineering should continue to arouse public apprehension the extremely desirable long-term medical goal of gene therapy may be threatened.

In answer to such a threat scientists have traditionally invoked the principle of freedom of inquiry, and in open societies they have interpreted this principle as a facet of the principle of freedom of expression. But perhaps it is no longer enough to wave the flag of Galileo: it is not inconceivable that as science digs deeper we might indeed acquire knowledge that is too hot to handle. However, if we accept this principle, and are therefore prepared to proscribed knowledge that clearly is predominantly dangerous, we must also recognize that no such knowledge has yet been demonstrated. And in the absence of a convincing demonstration, the guiding principle of free inquiry must still be defended. For virtually any basic knowledge about nature is ambivalent: it can be applied in both good and bad ways, and we have very limited capacity to foresee the full range of uses. We have even less capacity to foresee the social consequences of these uses (e.g., the automobile, television). The operational

conclusion, then, would be that we can best serve society's interests not by blocking knowledge itself but by being quicker to recognize specific harmful applications, and to prevent or to halt them.

Socially dangerous knowledge

The third concern, over scientific knowledge that shakes the foundations of public morality or the social order, began with Galileo, but it now seems restricted to biology—a field that impinges much more directly than physics on our views on human nature. In the 19th century it was the conservative establishment that saw a threat in Darwin's evidence against special creation; but today, ironically, it is the left that is most concerned with the possible social implications of a related kind of knowledge: that resulting from the modern fusion of evolutionary theory with genetics. There seem to be three fears: that research on genetic components of behavior will distract attention from the effort to provide social solutions to social problems; that the results of the research might conflict with egalitarian preconceptions about the amount and the distribution of variation in genetic potentialities in our species; and that the implications of the results may be distorted for political purposes.

The attack on this area of research would subordinate objective scientific knowledge to political dogma—a process that I have elsewhere (*Nature*, vol. 272, p. 390, 1978) called the moralistic fallacy, since it attempts to derive an 'is' from an 'ought' and hence is the mirror image of G. E. Moore's naturalistic fallacy. Moreover, the present attack presents a particularly close parallel to the frightening example of Lysenkoism. Nevertheless, this attack has evoked wide sympathy. The reasons are evident: our guilt over the inequities of slavery and race discrimination, and the all too real history of past premature extrapolations of evolutionary concepts by Social Darwinists, and past misuse of genetics by racists, to rationalize discriminatory practices. But in trying to prevent a repetition of this earlier subordination of science to reactionary political purposes, we must ask whether its present spiritual heirs are not those who would now subordinate science to progressive political purposes, rather than those who oppose its subordination to any politics, whether of the left or the right.

Hitler's ghost thus continues to haunt us, and to color our views on human genetics. And though in democratic countries we are unlikely to develop the overt mechanisms of Lysenkoism, with rigid governmental proscription of a politically offensive field, we must recognize that effective

suppression does not require this mechanism: in the face of the attacks of the last decade within universities, granting agencies in the United States are likely to be uncomfortable with such sensitive areas, and few graduate students are likely to enter the field. Indeed, because of such attacks several medical investigators have abandoned studies on the still poorly defined effects of an extra X or Y chromosome. Yet as was emphasized in the 1930s by J. B. S. Haldane, a British Marxist and leading geneticist, increased awareness and understanding of human genetic diversity offers great long-term promise for education (as well as more immediate promise for medicine).

The breadth of the issue, and the price of a victory for ideology over science, should not be underestimated. The problem will not remain confined to behavioral genetics: in time neurobiology, and evolutionary studies on social behavior, will surely also provide challenges to our preconceptions about human nature. Even today, efforts to inquire into biological and medical aspects of violence and crime, not necessarily genetic, encounter intense opposition from those who insist that the sickness lies entirely in society.

What is at stake is the intellectual freedom painfully acquired in part of the world over past centuries — not freedom to make or to do something potentially harmful, but freedom to know. The problem is one for all scholars, and not only for geneticists.

Of course, some would counter that it is callous to wish to unearth knowledge regardless of its political consequences. But we must recognize that the truths about human nature, both its universals and its diversity, will be there whether or not scientists discover them; and this reality will affect the success of those social policies that depend on assumptions about these matters. Moreover, if we recognize justice as a constantly evolving social construct it is difficult to see how any valid new knowledge can itself threaten justice. On the contrary, as we deepen our understanding of the interaction of inborn and social factors that influence human behavior we should be able to build more effective institutions of justice. Because of this possibility for positive uses, and also because we cannot unlearn the scientific method or eliminate human curiosity and creativity, it would appear more valuable for society to guard against the misuses of advances in genetics, rather than to seek protection by blocking off those advances.

Summary

On purely technical grounds it seems clear that recent fears of biology have been greatly exaggerated. In the area of dangerous products the history of the controversy over recombinant DNA research suggests that the scientific community did behave responsible, and that excessive involvement of the general public in the phase of technical assessment of risks impeded rather than advanced the process of arriving at reasonable judgments. In a second area, knowledge that may give us dangerous powers, restrictions on advances in the knowledge, as opposed to restrictions on its applications, are difficult to justify, simply because we are not able to foresee the full range of good and bad consequences of such advances.

A third concern, over increased insights into human nature and human diversity, is more complex. On the one hand, evidence that conflicts with treasured preconceptions may be painful. On the other hand, science gives us valuable insights as well as technological powers; and while the latter have generated the growing crisis of mankind, insight into the biological roots of our behavior may be what is most needed to meet that crisis. If we should cut off the flow of such insights, to avoid immediate problems, we may pay dearly in the long term—and we will also set a precedent that is inimical to an open, democratic society.

PROGRAMME

6th INTERNATIONAL CONGRESS OF LOGIC, METHODOLOGY AND PHILOSOPHY OF SCIENCE HANNOVER, AUGUST 22-AUGUST 29, 1979

Plenary Lectures

- R. W. FOGEL, "Scientific" History and Traditional History
W. HILDENBRAND, The Rôle of Mathematics in Economics
G. TAKEUTI, Work of Paul Bernays and Kurt Gödel
R. THOM, Role and Limits of Mathematization in Applied Sciences

Section 1: Proof Theory and Foundations of Mathematics

Invited Lectures

- J.-Y. GIRARD, A Survey of Π^1_1 -Logic
N. N. NEPEIVODA, Some Connections Between Proof Theory and Computer Programming
W. POHLERS, Admissibility in Proof Theory

Symposium Leader: R. GANDY

- The Role of Constructivity in Mathematics
M. HYLAND, Applications of Constructivity
P. MARTIN-LÖF, Constructive Mathematics and Computer Programming

Contributed Papers

- Y. GAUTHIER, Vers une Théorie de la Négation Locale
R. D. GUMB, The Craig Interpolation Lemma for (Free) Intuitionistic Logic with Equality
V. HARLIK, Syntactical Proofs of Interpolation for Fragments of $L(Q)$
N. NEPEIVODA, A Proof Theoretical Comparison of Program Synthesis and Program Verification
U. R. SCHMERL, Iterated Reflection Principles and the ω -Rule
H. SCHWICHTENBERG, Infinite Terms and Recursion Schemata
S. SHAPIRO, Pragmatic Properties and Mathematics
L. SZCZERBA, On "Geometrical Notions"
N. TENNANT, Proof Theory and Entailment
A. S. YESSENIN-VOLPIN, On an Explanation of an Anti-Traditional Paradox

Section 2: Model Theory and Its Applications

Invited Lectures

- Y. GUREVICH, Formal Topology
 A. MACINTYRE, Primes in Nonstandard Models of Arithmetic
 E. A. PALYUTIN, Completeness and Categoricity of Quasivarieties
 J. REINEKE, Algebraically Closed Commutative Rings

Contributed Papers

- H. ANDRÉKA, The Class of Neat-Reducts of Cylindric Algebras is Not a Variety
 Š. A. BASARAB, Remarks on a General Theory of Formally p -adic Fields
 A. M. DAWES, Games and Substructures
 I. GARRO, A Decision Procedure for p -adic Fields
 M. KRYNICKI, On Orderings of the Family of All Logics
 L. MAKSIMOVA, Almost All Modal Logics Do Not Have Craig's Interpolation Property
 Z. MIJAJLOVIC, Saturated Boolean Algebras
 I. MIKENBERG, R. CHUAQUI, Equational Classes of Partial Algebras
 R. MURAWSKI, Some Remarks on the Structure of Expansions
 A. OBERSCHELP, First Order Logic and Class Theory
 E. PALYUTIN, Criterion of Complete Varieties Categoricity
 M. PREŠIĆ, On the Embedding of Models
 S. PREŠIĆ, A Completeness Theorem for a Class of Propositional Calculi
 P. SCHMITT, Decidable and Undecidable Theories of Topological Abelian Groups
 G. TODT, A General Substitution Theorem
 I. TOTH, An Absolute Geometric Model of the Hyperbolic Plane and Some Related Metamathematical Consequences

Section 3: Recursion Theory and Theory of Computation

Invited Lectures

- A. N. DEGTEV, Small Degrees in Ordinary Recursion Theory
 D. NORMANN, Computations in Continuous Functionals
 V. R. PRATT, Dynamic Logic
 S. SIMPSON, Generalized Recursion Theory
 R. SOARE, Classical Recursion Theory

Contributed Papers

- ST. AANDERA, E. BÖRGER, Y. GUREVICH, Prefix Classes of Krom Formulae with Identity
 CH. HENNIX, A Finitistic Generalization of Recursiveness
 I. LAVROV, Computability of Partial Functions and Denumerability of Sets in Peano Arithmetic
 R. C. T. LEE, Applications of Symbolic Logic to Data Base Design

- S. C. LIU, Constructive and Non-Constructive Natural Numbers and Functions in Formal Number Theories
 R. PLIUŠKEVIČIUS, Some Syntactic Properties of Hoare-like Logic
 B. SCHINZEL, Some Results on Gödelnumberings and Friedbergnumberings

Section 4: Axiomatic Set Theory

Invited Lectures

- A. S. KECHRIS, Ordinal Games and Their Applications
 M. MAGIDOR, Compactness in the Constructible Universe
 D. A. MARTIN, On the Ordinal δ_6^1
 R. SOLOVAY, Rapidly Growing Ramsey Functions

Contributed Papers

- D. KUREPA, Models and Fix Points
 J. MLČEK, Valuations of Special Structures in Alternative Set Theory
 C. A. DI PRISCO, W. MAREK, Some σ -Algebras Containing Projective Sets
 M. v. RIMSCHA, A Rank-Function for Non-Founded Set Theory
 R. RUCKER, Class-Set Theories
 R. SAMI, Analytic Partitions of ${}^\omega\omega$ into Borel Sets of Bounded Rank
 A. SOCHOR, Elimination of Infinitely Small Quantities
 J. A. THOMAS, A New Foundation of Set Theory
 Z. VETULANI, Strong Constructibility for Arithmetics A_n
 A. S. YESSENIN-VOLPIN, On the Anti-Traditional Consistency Proof for ZF Set Theory

Section 5: Philosophical Logic

Invited Lectures

- J. McDOWELL, Truth-Value Gaps
 J. M. DUNN, A Sieve for Entailments
 E. K. VOISHVILLO, Semantics of General State Descriptions
 R. WÓJCICKI, Referential Matrix Semantics for Propositional Calculi

Contributed Papers

- E. M. BARTH, A Normative Foundation for First-Order Logic, Based on the Completeness of the Dialogue Method
 L. M. BATTEN, D. N. WALTON, Graphs of Arguments
 J. VAN BENTHEM, Partial Logical Consequence
 M. BEZHANISHVILI, On the Operative Understanding of Negation
 S. BUGAJSKI, A Logical Structure Arising in Operational Quantum Mechanics
 M. L. DALLA CHIARA, A. M. PAOLINI, Completeness Properties of Non Distributive Logics
 N. E. CHRISTENSEN, Logical Constants and Their Context

- W. S. CRODDY, Necessity and Singular Terms
 P. GÄRDENFORS, An Epistemic Approach to Conditionals
 K. GALÁNTAI HAVAS, Inconsistent Formal Systems and Dialectics
 C. F. GETHMANN, Zur Pragmatik des konstruktiven Subjunktors
 S. GUCCIONE, S. TERMINI, The Modal "Vaguely"
 L. GULLVÅG, The Logic of Thinking and Discrimination
 R. HEGSELMANN, Zum Program einer Logik der Argumentation
 W. HEITSCH, Eine logizistische Begründung der Modalitätentheorie im Sinne Frege's
 J. HUMPHRIES, Truth and Ontological Commitment
 C. JOJA, Le principe d'analogie et la rationalité moderne
 R. KAUPPI, Zur Logik der Veränderung
 W. KINTD, Die Logik von Sprachen mit unfundierten Formeln
 V. KOSTIOUK, Propositional Epistemic Logic with Sublogics of Provability and Empirical Acceptance
 W. LENZEN, The Logic of "Is Probable"
 B. LOEWER, Partial Order vs. Total Order Semantics for Counterfactuals
 A. MENNE, Zum Anwendungsproblem der Logik
 I. NIINILUOTO, Degrees of Truthlikeness: Singular Sentences vs. Generalizations
 CH. PARSONS, The Logic of Sense and Denotation and Montague's Intensional Logic
 C. POPA, O. STĂNĂȘILĂ, Norm Systems and Valuation of Normed Actions
 J. SANCHEZ, Intuitive Semantics
 O. SEREBRYANNIKOV, New Cut-Free Calculi of Sequents in Modal Logic
 V. SMIRNOV, Logical Systems with Modal Temporal Operations
 E. SMIRNOVA, An Approach to the Semantics of Non-Extensional Contexts
 W. SPOHN, A Definition of "Cause"
 G. STAHL, Momentary and Temporally Extended Individuals
 V. TSELISHCHEV, Essentialism, Singular Terms and De Re and De Dicto
 D. VANDERVEKEN, What is an Illocutionary Force?
 H. WESSEL, Widersprüchliche Theorien und logische Folgebeziehung
 J. H. WOODS, D. N. WALTON, Formal Logic and the Logic of Argument

Section 6: General Methodology of Science

Invited Lectures

- R. BHASKAR, Realism in the Natural Sciences
 V. N. KOSTIOUK, Possible Worlds and Ontology of Scientific Theories
 J. J. C. SMART, Difficulties for Realism in the Philosophy of Science
 P. SZTOMPKA, Metatheoretical Dilemmas of the Social Sciences. The Case of Sociology

Contributed Papers

- R. ABEL, "What is an Explanandum" ?
 M. A. AMER, Towards a Philosophy of Mathematics Based on Its Relationship to Reality
 L. BAZHENOV, Is Mathematization an Obstacle for Cognition of the Qualitative Peculiarity of Reality?
 W. BOOS, Model-Theoretic Realism

- J. VAN BRAKEL, H. VERMEEREN, On the Philosophy and the Foundations of Chemistry
- B. CHENDOV, On the Complex Theories: A Specific Type of Synthesis of Scientific Problems
- G. J. DALENOORT, A General Definition of the Concepts of Force, Inertia and Resistance, Applicable for Metaphorical Use
- G. DORN, The Method of Explication
- T. FÖLDESI, Über das Problem des Wahrseins der Fragen und Normen
- H. FRANKEL, Why Harry Hess's Seafloor Spreading Hypothesis Was Accepted with the Confirmation of the Vine-Matthews-Hypothesis
- D. GORSKI, Identifizierung des Unidentifizierbaren
- V. S. GOTTF, General Scientific Nature of Mathematical Concepts
- L. HÄRSING, Ein neuer Versuch, das Prinzip des ausgeschlossenen Widerpruchs und das dialektische Widerspruchsprinzip in Einklang zu bringen
- W. HEITSCH, Zur Funktion der Frage in der wissenschaftlichen Forschung
- H. S. JENSEN, Marx, Mathematics and Materialism
- W. B. JONES, The Structure of Theories Revisited
- W. KLEVER, Terminal Objectivity
- G. KLIMASZEWSKY, Zur Aktivität des Subjekts in der wissenschaftlichen Erkenntnis
- C. R. KORDIG, Self-Reference and the Philosophy of Science
- V. KOURAEV, Le sens et la valeur de la formalisation dans la science
- G. KOVACS, Синтез наук и методологическая функция философии
- L. KRÜGER, History and Philosophy of Science: A Marriage for the Sake of Reason
- M. KÜTTNER, A Solution to Ruling out All Partial and Total Self-Explanations in D-N Arguments
- TH. A. F. KUIPERS, Diminishing Returns from Repeated Tests
- W. I. KUPZOW, Über die Korrelation genetisch miteinander verbundener wissenschaftlicher Theorien
- J. MAKSABEDIAN, The Experimental Method in Technology
- C. MARE, The Role Played by Mathematics Within the Framework of Interdisciplinary Communications
- E. McMULLIN, The Language of the Book of Nature
- I. MERKULOV, The Justification Problem of Science and Hypothetical Method
- I. NOVIK, Heuristic Value of the Global Simulation
- A. PEDERSEN, J. WITT-HANSEN, Futures Research, Its Subject-Matter and Methods
- C. POPA, Semantic Definition of the Concept of Method
- M. POPOVICH, Notes About Analysis and Synthesis
- D. SCHULZE, Methodologische Probleme der Theorie der unscharfen Mengen
- C. SENECA, U. KNAUER, Historical vs. Formal Approach
- V. STOLJAROV, Content-Genetic Investigation of Cognition in Logics and Methodology of Science
- R. TUOMELA, Analogy and Distance
- J. B. UBBINK, J. VAN BRAKEL, Observation Sentences and the Concept of Truth
- CH. C. VERHAREN, The Mediation of Methodological Disputes in Philosophy of Science
- I. ZAPLETAL, The Methodological Role of the Mathematization of Scientific Cognition

Section 7: Foundations of Probability and Induction*Invited Lectures*

- B. DE FINETTI, Probability: The Different Views and Terminologies in a Critical Analysis
 read by I. LEVI
- T. SEIDENFELD, Paradoxes of Conglomerability and Fiducial Inference
- M. STONE, Review and Analysis of Inconsistencies Related to Improper Priors and Finite Additivity

Contributed Papers

- TH. A. BRODY, The Ensemble Conception of Probability
- R. CHUAQUI, Probability as Between Truth and Falsehood
- L. J. COHEN, Bayesianism Versus Baconianism in the Evaluation of Medical Diagnoses
- C. COSTANTINI, A Rational Reconstruction of the Beta Distribution
- A. DALE, A Finite Form of De Finetti's Theorem for a First Order Language
- T. FINE, Some Possibilities for Modal and Comparative Probability Concepts
- P. KRAUSSER, Grue Once More
- J. ŁOŚ, Troubles with Events for an Invariant Probability
- J. VON PLATO, On a Reductive Relation in the Interpretation of Probability

Section 8: Foundations and Philosophy of the Physical Sciences*Invited Lectures*

- E. BELLONE, The Second Scientific Revolution and the Interaction between Mathematics and Theoretical Physics; Fourier, Hamilton and Boltzmann
- A. P. GRECOS, The Problem of Irreversibility in Theoretical Physics
- Y. SACHKOV, Probability in Classical and Quantum Physics
- D. SHAPERE, The Scope and Limits of Scientific Change

Contributed Papers

- I. AKCHURIN, New Type of Complementarity in Physics — Complementarity of Topology and Logic
- J. ALMOG, A Physical Metamorphosis in Logic?
- L. ANTIPENKO, On the Limits of Recursive Axiomatization of Quantum Mechanics
- R. BORN, Causality Versus Quantumlogic?
- J. VAN BRAKEL, A. A. VAN DE PEUT, Alternative Physical Concatenation Procedures in Length Measurements
- S. BUGAJSKI, A New Concept of Quantum Mixed States
- M. GHINS, Géométrie et expérience: Remarques sur le conventionalisme de Henri Poincaré
- R. GILES, A Pragmatic Approach to the Foundations of Physics
- A. GRIEDER, Relativity and the Question of Preferred Reference Systems
- R. INHETVEEN, Geometrien und Modelle
- A. KAMLAH, Uniqueness of Space-Time Measurement due to Invariance Principle of Physics
- P. KROES, On the Geometrical Nature of Physical Time
- U. MAJER, Das Verhältnis von Objekt- und Meßtheorie als methodische Relation

- K. MANDERS, What is Quantitative about Physical Measurement?
- E. MARQUIT, Reversible Ultimate Causes and Development in Physics
- P. MITTELSTAEDT, Modalities in Quantum Logic
- A. PECHENKIN, Foundations of Physics: Mathematics Versus Natural Philosophy
- J. PINKAVA, Application of the Correlation Method in Chemistry
- H. A. PUYAU, Sur le conventionnalisme et l'apriorisme en géométrie appliquée
- C. RAJSKI, On Logic of Non-Individuals
- M. REDHEAD, Experimental Tests of the Sum Rule
- U. RÖSEBERG, Unbestimmtheit, Komplementarität, Widerspruch
- L. SOFONEA, L'idée de "deux" dans la pensée de la physique
- E. W. STACHOW, Operational Quantum Probabilities
- G. STAVENGA, The Subject-Object Relation in Relativity Theory and in Quantum Theory, and the Development of Theoretical Physics
- P. VAN DER VET, The Law of Mass Action, the Nernst Equation, and Isotopes
- G. VOLLMER, Objektivität und Invarianz
- E. W. WETTE, Universality of the Geometro-Static Representation of All Motions

Section 9: Foundations and Philosophy of Biology

Invited Lectures

- L. DARDEN, Aspects of Theory Construction in Biology
- O. REIG, The Ontological Nature of Biological Species
- A. WOODFIELD, Some Connections Between Ascriptions of Goals and Assumptions of Adaptiveness

Contributed Papers

- J. ALMOG, A Quantum Basis of Heredity and Mitosis?
- J. BEATTY, Optimal Design Models, and the Strategy of Model Building in Evolutionary Biology
- A. LINDEMAYER, N. SIMON, The Problem of Theory Reduction in Genetics
- M. B. WILLIAMS, Circularity at the Core of a Theory: Deep Tautology Versus Mere Tautology

Section 10: Foundations and Philosophy of Psychology

Invited Lectures

- N. J. BLOCK, Psychologism Vindicated
- K. CAMPBELL, Land's Theory of Colour Vision and Its Philosophical Implications
- D. FØLLESDAL, Intentionality and Behaviorism
- B. VELICHKOVSKY, Actual Problems of Cognitive Psychology

Contributed Papers

- CH. CHIANG, Uncertainty Principle in Physics and Illusory Perception in Psychology
- G. J. DALENOORT, Deterministic Versus Teleological Explanation in General Systems Theory
- M. DASCAL, On the "Expressive Power" of the "Language of Thought"
- G. HELLMANN, Physicalism and the Mind-Body Problem

B. M. KEDROV, Three Units in the Analysis of Scientific and Technological Creative Work (Psychology, History, Logic)

W. K. WANG, The Limitation of Human Knowledge

R. WERTH, Does an Observational-Theoretical Split Make Sense in Psychology?

Section 11: Foundations and Philosophy of the Social Sciences

Symposium

Equilibrium Economics

E. MALINVAUD, The Equilibrium Concept in Economics

Commentator: P. HAMMOND

Symposium

Formal Systems of Rights

P. HAMMOND, Liberalism, Independent Rights and the Pareto Principle

Commentator: A. GIBBARD

Contributed Papers

D. CLARKE, Assumptions of Rationality in Social Science Explanations

O. GRUENGARD, Mathematical Toys in the Behavioral Sciences

J. N. KAUFMANN, Explication par des lois et explication par des règles

V. KELLE, Methodological Problems of Historical Analysis of the Interconnection Between Science and Society

R. LUESCHER, Anti-Reductionist Name-Calling

S. MAFFETTONE, Epistemology and Hermeneutics

E. MARKARIAN, The Methodological Principles of Studying the Local Diversity of Culture

R. MATTESSICH, Knowledge and Utility: Structural Interrelations from a Systems Point of View

R. MIGUELEZ, Connaissances et actions sociales

C. SAVARY, La pragmatique et le concept d'idéologie

D. M. DE SOUZA, Linguistic Philosophy as a Critical Analysis

A. STARCHENKO, "Conviction": A Functional Analysis

Section 12: Foundations and Philosophy of Linguistics

Invited Lectures

Z. HARRIS, Mathematical Analysis of Language

H. HEIDRICH, Formal Capacity of Montague Grammars

R. MARTIN, On Logico-Linguistics; Structure, Transformation and Paraphrase

D. SANKOFF, Sociolinguistic Method and Linguistic Theory

Contributed Papers

M. H. DASCAL, Some Questions About Functionalism

J. HINTIKKA, On Any-Thesis and the Methodology of Linguistics

S. LEBLANC, Pour une caractérisation des contextes d'emploi d'énoncés

A. TER MEULEN, Constraints on Logical Form

- N. MOUTAFAKIS, The Mathematical Foundations of Rescher's Logic of Preference
 J. PERE, "Regularization: Formal and Functional"
 PH. L. PETERSON, On the Natural Logic of Complex Event Expressions
 J. POGONOWSKI, Tolerances in Linguistics
 J. D. RINGEN, "Quine on Introspection in Linguistics"
 E. SAARINEN, Meaning and Use: Methodological Remarks on Game-Theoretical Pragmatics
 R. ZUBER, On Metalanguage in Language

Section 13: History of Logic, Methodology and Philosophy of Science

Invited Lectures

- M. A. FINOCCHIARO, Methodological Problems in the History of Science: An Analytical Approach
 V. LEKTORSKY, On the Change in the Relationship Between Science and Epistemology of Science as Concerns Historical Development
 N. RESCHER, Kant on Scientific Questions
 S. SURMA, On the Origin and Subsequent Applications of the Concept of Lindenbaum Algebra

Frege Symposium Leader: H. HERMES

- CH. THIEL, From Leibniz to Frege: Mathematical Logic between 1679 and 1879
 I. ANGELELLI, Frege's Notion of Bedeutung

Contributed Papers

- P. ACZEL, The Structure of the Formal Language of Frege's Grundgesetze
 F. ALTRICHTER, Cognosco venientem
 H. BARREAU, Une nouvelle et double reconstitution du maître-argument de Diodore Cronos
 K. BERKA, Bolzano's Lehre von den Wahrscheinlichkeitsschlüssen
 A. BOBOC, Husserl and the Programme of Modern Logic
 J. W. DAVIS, Hume's Account of Probability
 R. R. DIPERT, Boolean Propositional Logics
 C. EISELE, Mathematical Methodology in the Foundation of Peirce's Thought
 A. GRIGORIAN, V. KIRSANOV, Mathematics and Evolution of Classical Mechanics
 I. HRONSZKY, Methodenentwicklung und methodologische Reflexionen in den experimentellen Wissenschaften in ihrer "vorparadigmatischen" Phase (am Beispiel der Chemie)
 S. KNUUTTILA, The Change of Modal Paradigms in Late Medieval Philosophy
 P. KRAUSSER, D. KRAUSSER, Infinities. The Thesis of Kant's First Antinomy
 C. A. LERTORA-MENDOZA, ROBERT GROSSETESTE: Astronomie et astrologie au début du XIII^e s.
 G. LEVIN, The Medieval and Modern Interpretation of the Problem of Universalias
 A. MADARÁSZ, Frege and Modern Intensional Logic
 L. MARKOVA, Problems of the Choice in the Historiography of Science

- B. MOSS, Survival or Resurrection? The Recent History of Infinitesimals
 J. M. B. MOSS, How Can Frege's Logical Objects be Known?
 V. MUÑOZ-DELGADO, Logic Humanism and Science in Salamanca (1490-1554)
 M. H. OTERO, Frege vs. Hilbert; Revolutionary Changes in Geometry
 C. PANACCIO, "Suppositio Naturalis" au XIII^e siècle et signification chez Guillaume d'Occam
 I. PÂRVU, Riemann vs. Kant: A Case Study For a New Historiography of the Philosophy of Science
 E. PICARDI, A Note on Frege's Context Principle
 V. SADOVSKY, The Paradoxes of the Systems Thinking and the History of the Methodology of Science in the XXth Century
 M. SANCHEZ-MAZAS, Un modèle arithmétique de la syllogistique et ses extensions
 G. SCHUBRING, Die Konzeption der Reinen Mathematik und ihre Bedeutung für eine Anwendungsorientierung der Mathematik
 H. SHEEHAN, The History of the Philosophy of Science: A Broader Perspective
 K. SUNDARAM, Falsificationism and Research Programs: A Case Study in Chemistry
 I. TOTH, The Genetic Structure of the History of Non-Euclidean Geometry and Some Related Theoretical Problems
 N. YULINA, The History of Metaphysics in the XXth Century and the Images of Science
 SH. ZELLWEGER, A Logical Garnet as Both a 3-D and a 4-D Symmetry Model of the 16 Binary Connectives

Section 14: Fundamental Principles of the Ethics of Science

Symposium

- Distributive Justice and the Allotment of Resources of a Society to Scientific Research
 P. SUPPES, Rational Allocation of Resources to Scientific Research
 Commentator: B. HANSSON

Symposium

- Ethical Problems Involved in Gene Research and Manipulation
 S. P. STICH, On Genetic Engineering, the Epistemology of Risk, and the Value of Life
 Commentator: B. D. DAVIS

Contributed Papers

- J. W. R. FENNEMA, On Foundations and Fundamental Principles
 P. H. ROSSEL, Science-Based Paternalism in the Medical Ethics of Th. Percival
 K. E. TRANØY, Rationality of Science: Consensus or Certainty?

INDEX OF NAMES

- Ackermann 80
Anscombe 343
Archimedes 4
Aristotle 479, 531, 755, 756
Arrow, K. 607
Aumann, R. 607
Ax, J. 197, 198, 203

Bahcall, J. 449
Barcan-Marcus, R. 363
Bardili, Ch. G. 767
Bateson 482
Beard, Ch. A. 19, 22, 44
Becker, C. 19
Beecher Stowe, H. 37
Benacerrof, P. 85
Benado, M. 497
Bennett, M. 639, 640
Berkeley, B. 143
Bernays, P. 77, 78, 79
Bernoulli, J. 759, 760, 763
Berry, P. 497
Bishop, E. 142
Bloch, M. 22
Bocheński, J. M. 767
Boffa 193
Bohr, N. 712
Boje, P. 15
Bolzano, B. 760
Boole, G. 755, 756, 764, 765
Borel, E. 393
Born 434
Bourbaki 657
Bowley 68
Boyd, R. 360
Braudel, F. 22, 23
Brennan, G. 607
Brentano 554
Bridgeman 712
Brouwer 79 141, 143, 770

Butlin, N. 15
Butterfield, H. 22

Cantor, G. 141
Carnap, R. 398
Carson 227
Cartwright, N. 363
Castillon, F. von 762, 767
Castillon, J. de 767
Cauchy, A. L. 4
Cavalieri 4
Chang, C. C. 287
Church, A. 767
Clark, M. 291
Clarke, S. 365
Clive, J. 15
Coats, A. W. 15
Cocchiarella 639
Cohen, P. 84
Cole, D. 363
Collingwood 18
Cooper 639, 650
Couturat, L. 760, 762
Croce, N. 18
Cromwell, Th. 44

Darwin, Ch. 17, 467, 472, 840
De Champeaux, G. 482
De Finetti, B. 396, 398
Dekker 237
De Morgan, A. 762, 764
Descartes, R. 712, 714
Dilthey 18
Dirac, P. A. 4
Dries, van den 193, 197
Drobisch, M. W. 763
Dummett, M. 369
Duns Scotus 480
Dvornickov, S. 243

- Einstein, A. ..., 141, 360, 658, 712
 Engerman, St. R. 15
 Etchemendy, J. 363
 Faraday, M. 452
 Fatou 12
 Febvre, L. 22
 Feferman, S. 79
 Feyerabend, P. 337, 346, 348, 694
 Fischer, D. H. 15
 Fogel, E. G. 52
 Fohlen, C. 15
 Ford 833
 Fourier, J. 4
 Freeman, E. A. 21
 Frege, G. 160, 299, 301, 305, 755, 756,
 757, 765, 766, 768, 769, 770
 Friedberg 266, 268
 Friedman, G. 15
 Gabbay 639
 Gadamer 566
 Galenson, D. 15
 Galilei, G. ..., 712, 839, 840
 Gallin 639
 Gandy, R. 242
 Gauss, C. F. 455
 Geach 291
 Gentzen 79, 82, 83, 124, 125, 128
 Gergonne, J. D. 763
 Gibbon 46
 Gödel, K. 77, 78, 79, 80, 81, 82, 83, 84,
 85, 673, 770
 Grassmann, R. 756, 760
 Grzegorczyk, A. 90
 Haack, S. 363
 Hacking, I. 363
 Hahn, F. 607
 Haldane, J. B. S. 841
 Hamaker 394
 Hamblin 639, 762
 Hamilton, S. W. 762, 764
 Hammond, P. J. 597
 Hegel, G. F. W. 716, 763
 Heijenoort, J. van 755, 756, 757
 Heisenberg 4, 712
 Helmholtz 576
 Hénon 12
 Herlihy, D. 15
 Hesse, M. 694
 Hilbert, D. 77, 78, 79, 80, 82, 84, 125,
 141, 770
 Hintikka, J. 639
 Hirschmann, D. 501
 Hobsbawm, E. 15
 Holland, G. J. von 762
 Hoorman Jr., C. F. A. 762
 Hughes, Ch. E. 20
 Humboldt, W. von 18
 Hume, D. 364, 365
 Hunter, G. 291
 Husserl, E. 554, 555, 559
 Hyland, J. M. E. 142, 242
 Jackson, R. H. 20
 Jastrow 554
 Jeffreys 413
 Jensen, U. J. 265, 267
 Jevons 63
 Jourdain, Ph. E. B. 756
 Julia 12
 Jungius, J. 759
 Kamp 639
 Kant, I. 357, 716, 763, 767, 798
 Kaplan 639
 Karttunen 639
 Keats, J. 28
 Keller, M. 15
 Kepler, J. ..., 455
 Kleene, S. C. 83, 118, 130, 265, 266
 Kneale, M. 755
 Kneale, W. 755
 Kolmogoroff, A. N. 13
 Kowalewska, S. W. 4
 Koyré 337
 Kreisel, G. 79
 Kripke, S. 321, 639
 Kuklick, B. 363
 Kulikoff, A. 15
 Lakatos 337, 340
 Lambert, J. H. 762, 763

- Land, E. H. 541
 Laplace 454
 Larman, D. G. 424
 Lane, F. C. 15
 Lavoisier 453
 Leibniz, G. W. 5, 365, 455, 714, 755,
 756, 757, 758, 759, 760, 762, 763, 764,
 765, 767
 Lenin, V. I. 357
 Lerman 266, 268
 Leśniewski, St. 658, 719, 768
 Lewis, C. I. 762, 764, 767
 Lindley, D. V. 413
 Linnaeus, C. 483
 Locke, J. 19, 364
 Lorenz 12
 Lorenzen, P. 755
 Łoś, J. 15, 607
 Lotze, R. H. 763
 Lukasiewicz, J. 719, 758
 Lullus, R. 764
 Maass 265, 266, 268
 Macaulay, lord 21
 Maehara 79
 Malthus 16
 Markov, A. A. 109
 Marshall 18, 63
 Martin-Löf 142, 143
 Marx, K. 63, 586
 Maskin, E. 607
 Maslow 13
 Maxwell, J. C. 442
 McAloon 193
 McColl, H. 756
 McMullin 694
 Menger 63
 Michelet, J. 21
 Mill, J. S. 63, 604
 Miller, D. 370
 Millikan 367
 Mises, R. von 393
 Mittelstrass, J. 766
 Montague, R. 639, 640, 641
 Moore, G. E. 840
 Morgan 16
 Morley, M. 214, 216
 Moschovakis, Y. 265
 Mostowski, A. x
 Muchnik 266, 268
 Naess, A. 555
 Nagel 348
 Namier, L. 44
 Nash, J. 12
 Nepeivoda, N. N. 142
 Neumann, J. von 11
 Nevins, A. 22, 24
 Newton, I. 8, 360, 361, 454, 455, 712, 713
 Niiniluoto, J. 356
 Ockham, W. of 481
 Orwell, G. 365
 Pacholski, L. 193
 Papst, W. 756
 Pareto 63, 607
 Parsons 639, 650
 Partee, B. 639, 641, 652
 Pascal, B. 4
 Peano, G. 748, 770
 Peirce, C. S. 764, 765
 Pestel, E., ix
 Pettengill, J. 607
 Planck, M. 43
 Plato 479
 Play, Le 16
 Ploucquet, G. 762, 763
 Plumb, G. H. 22
 Poincaré, H. 432
 Poizat 193
 Polya, G. 141
 Pontrjagin, L. S. 13
 Popper, K. 337, 370, 711
 Prawitz, D. 79
 Prescott 46
 Putnam, H. 85
 Quine, W. 340, 358, 360
 Ranke, L. von 17, 21, 28
 Raspe, R. E. 760, 762
 Rauszer, C. 193
 Ray, J. 481

- Redi 481
 Reichenbach 367
 Ricardo, D. 63
 Riemann 455
 Risse, W. 767
 Rodman 639
 Roosevelt, F. D. 24, 25, 47
 Rorty, R. 449
 Rosenkrantz, B. 15
 Rosser 269
 Royce, J. 761
 Russell, B. 78, 83, 84, 657, 667, 752, 768
 Saccheri, G. 757, 760, 761, 762
 Samuelson 66
 Sanches, H. 319
 Santilli, E. 497
 Schmidt, K.-D. 574
 Scholz, H. 761
 Schröder, E. 756, 764, 765
 Schrödinger 4
 Schütte 79
 Scott 639
 Sechenov 577
 Segner, J. A. von 762
 Sellars, W. 355, 356
 Shakespeare, W. 46
 Shapley 591
 Shearmen, A. T. 767
 Shore, R. A. 264, 266
 Sierpiński, W. 719
 Silver, J. 268
 Skolem, Th. 80
 Smart, J. 356
 Smiley, T. J. 291
 Snyderman, N. 449
 Soare 267
 Sokoloff, K. 15
 Sonnenschein 74
 Spielmann, S. 402
 Stalnacker, R. 363
 Stone, M. 396, 397, 401, 403, 405, 407, 411
 Stout 548
 Styazhkin, N. J. 759, 767
 Swammerdam 481
 Szpilrajn-Marczewski, E. 617, 720
 Tait 242
 Tarski, A. 143, 368, 369
 Tawney, R. H. 22, 24
 Thomas Aquino 364, 365, 480
 Thomson, J. J. 367
 Thomason, R. 639
 Thucydides 52
 Tichy, P. 370
 Tilly 41
 Toulmin 338
 Trendelenburg, F. A. 763, 764
 Trevelyan, G. M. 22
 Truman, H. S. 20
 Tumela, R. 356
 Turner, F. J. 22
 Tweten, A. D. C. 763, 767
 Urquhart, A. 291
 Venn, J. 762, 767
 Vico, G. 21, 337
 Victorin, A. 767
 Vinson, F. M. 20
 Vries, de 482
 Wagner, G. 760
 Washington, G. 29
 Watson, J. 559, 820
 Weber, M. 802
 Weyl, H. 79, 141
 White, H. 15
 White, M. 449
 Whitehead, A. N. 78, 83, 84, 377
 Wilson, M. 363
 Wittgenstein, L. 554
 Womack, J. 15
 Wright, von 291, 343
 Yukawa 456
 Zermelo, E. 79, 432, 657