

STUDIES IN LOGIC
AND
THE FOUNDATIONS OF MATHEMATICS

VOLUME 126

J. BARWISE / H.J. KEISLER / P. SUPPES / A.S. TROELSTRA
EDITORS

***Logic, Methodology
and
Philosophy of Science
VIII***

Edited by

J.E. FENSTAD, I.T. FROLOV and R. HILPINEN

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD • TOKYO

LOGIC, METHODOLOGY AND PHILOSOPHY OF SCIENCE VIII

STUDIES IN LOGIC
AND
THE FOUNDATIONS OF MATHEMATICS
VOLUME 126

Editors

J. BARWISE, *Stanford*
H.J. KEISLER, *Madison*
P. SUPPES, *Stanford*
A.S. TROELSTRA, *Amsterdam*

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD • TOKYO

LOGIC, METHODOLOGY AND PHILOSOPHY OF SCIENCE VIII

PROCEEDINGS OF THE EIGHTH INTERNATIONAL
CONGRESS OF LOGIC, METHODOLOGY
AND PHILOSOPHY OF SCIENCE,
MOSCOW, 1987

Edited by

Jens Erik FENSTAD

University of Oslo, Oslo

Ivan T. FROLOV

Academy of Sciences of the USSR, Moscow

Risto HILPINEN

University of Turku, Turku

University of Miami, Coral Gables



1989

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD • TOKYO

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

Sole Distributors for the U.S.A. and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
655 Avenue of the Americas
New York, NY 10010, U.S.A.

Library of Congress Cataloging-in-Publication Data

International Congress of Logic, Methodology, and Philosophy of
Science (8th : 1987 : Moscow, R.S.F.S.R.)

Logic, methodology and philosophy of science VIII : proceedings of
the Eighth International Congress of Logic, Methodology and
Philosophy of Science, Moscow, 1987 / edited by Jens Erik Fenstad,
Ivan T. Frolov, Risto Hilpinen.

p. cm. -- (Studies in logic and the foundations of
mathematics ; v. 126)

Includes bibliographies and index.

ISBN 0-444-70520-1

1. Science--Philosophy--Congresses. 2. Science--Methodology--
Congresses. 3. Logic--Congresses. 4. Mathematics--Philosophy--
Congresses. I. Fenstad, Jens Erik. II. Frolov, Ivan Timofeevich.
III. Hilpinen, Risto. IV. Title. V. Series.

Q174.I58 1987

501--dc19

88-39426

CIP

ISBN: 0 444 70520 1

© Elsevier Science Publishers B.V., 1989

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 written permission of the publisher, Elsevier Science Publishers B.V. (North-Holland), P.O. Box 103, 1000 AC Amsterdam, The Netherlands.

Special regulations for readers in the USA – This publication has been registered with the Copyright Clearance Center Inc. (CCC), Salem, Massachusetts. Information can be obtained from the CCC about conditions under which photocopies of parts of this publication may be made in the USA. All other copyright questions, including photocopying outside of the USA, should be referred to the publisher.

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein.

Printed in The Netherlands

PREFACE

This volume constitutes the Proceedings of the Eighth International Congress of Logic, Methodology and Philosophy of Science. The Congress was held at Moscow University, USSR, from August 17 to August 22, 1987, under the auspices of the Division of Logic, Methodology and Philosophy of Science of the International Union of History and Philosophy of Science. The Congress was sponsored by the Academy of Sciences of the USSR. It was organized by its Local Organizing Committee in close cooperation with its Programme Committee and the Executive Committee of the Division of Logic, Methodology and Philosophy of Science. The scientific programme of the Congress was drawn up by the Programme Committee together with 13 Advisory Committees which corresponded to the 13 Sections of the Congress (Sectional Committees). (A list of the members of the various committees is appended to this Preface.) The 13 Sections of the Congress were as follows:

1. Foundations of mathematical reasoning
2. Model theory
3. Foundations of computing and recursion theory
4. Set theory
5. General logic
6. General methodology of science
7. Foundations of probability and statistical inference
8. Foundations of physical sciences
9. Foundations of biological sciences
10. Foundations of psychology and cognitive sciences
11. Foundations of social sciences
12. Foundations of linguistics
13. History of logic, methodology and philosophy of science

Each Section comprised a few invited addresses as well as a large number of contributed papers. In addition to the Sections, the programme included the Inaugural Address by Academician V.N. Fedoseyev and two Intersectional Symposia, "New Patterns of Explanation in Science" and "Science and Ethics". This volume contains only the invited

addresses of the Congress; a list of the contributed papers is given at the end of the volume.

We should like to thank the authors and Elsevier Science Publishers B.V. for their support to our editorial work, Dr. Tore Langholm of the University of Oslo, and Mrs Terttu Bylina, Mrs Riitta Lehtikoinen and Mrs Rita Luoma of the University of Turku for editorial assistance.

Oslo, Moscow and Coral Gables
March 1988

J.E. FENSTAD
I.T. FROLOV
R. HILPINEN

APPENDIX TO THE PREFACE

List of the members of the Executive Committee of the Division of Logic, Methodology and Philosophy of Science, International Union of History and Philosophy of Science in 1987:

Dana S. Scott	USA	President
Paul Weingartner	Austria	1st Vice-President
Andras Hajnal	Hungary	2nd Vice-President
Risto Hilpinen	Finland	Secretary
Helmut Pfeiffer	FRG	Treasurer
Jerzy Łoś	Poland	Past-President

List of the members of the Programme Committee of the 8th International Congress of Logic, Methodology and Philosophy of Science:

Jens Erik Fenstad	Norway	Chairman
Justus Diller	FRG	
Yuri L. Ershov	USSR	
Max Jammer	Israel	
Francisco Miro Quesada C.	Peru	
Wesley C. Salmon	USA	

List of the members of the Sectional Programme Committees:

Section 1:

H. Schwichtenberg	FRG	Chairman
G. Takeuti	USA	
S.I. Adjan	USSR	

Section 2:

L. Pacholski	Poland	Chairman
R.L. Vaught	USA	
G. Sabbagh	France	

Section 3:

J. Shepherdson	England	Chairman
R.I. Soare	USA	
A.N. Degtev	USSR	

Section 4:

A. Hajnal	Hungary	Chairman
M. Magidor	Israel	
C.A. Di Prisco	Venezuela	

Section 5:

J. van Benthem	The Netherlands	Chairman
M. Dalla Chiara	Italy	
G. Boolos	USA	

Section 6:

J. Hintikka	Finland	Chairman
D.P. Gorsky	USSR	
K. Berka	Czechoslovakia	

Section 7:

D.H. Mellor	England	Chairman
R. Chuaqui	Chile	
I. Niiniluoto	Finland	

Section 8:

C. Hooker	Australia	Chairman
R. Toretto	Puerto Rico	
M.M. Yanase	Japan	

Section 9:

D. Hull	USA	Chairman
J. Hodge	England	
R.S. Karpinskaya	USSR	

Section 10:

E.N. Sokolov	USSR	Chairman
H. Rouanet	France	
P. Suppes	USA	

Section 11:

P. Gärdenfors	Sweden	Chairman
A. Gibbard	USA	
W. Hildenbrand	FRG	

Section 12:

M. Bierwisch	GDR	Chairman
O. Dahl	Sweden	
R. Cooper	Sweden	

Section 13:

V.A. Lektorsky	USSR	Chairman
R. Butts	Canada	
C. Thiel	FRG	

List of the members of the Local Organizing Committee:

Ivan T. Frolov Chairman

D.P. Gorsky Vice-Chairmen

Yu.L. Ershov

V.I. Kuptsov

V.A. Lektorsky

S.T. Melyukhin

Yu.V. Sachkov

V.S. Stepin

I.S. Melyukhin Secretaries

S.A. Nikolsky

S.I. Adyan

G.G. Chakhmakhchev

A.G. Egorov

Yu.S. Eliseeva

V.I. Fatina

L.Ja. Gervits

V.S. Gott

D.P. Gribanov

R.S. Karpinskaya

V.J. Kelle

A.D. Kosichev

G.G. Kvasov

N.I. Lapin

I.A. Lavrov

B.F. Lomov

N.N. Moiseev

P.S. Oraevsky

V.V. Petrov

A.I. Rakitov

V.N. Sadovsky

V.S. Semenov

V.I. Shinkaruk

E.A. Sidorenko

V.A. Smirnov

E.N. Sokolov

S.V. Yablonsky

R.G. Yanovsky

This Page Intentionally Left Blank

CONTENTS

Preface	v
Appendix to the Preface	vii
Presidential Address, <i>D.S. Scott</i>	xv

INAUGURAL ADDRESS

Philosophy, Science and Man, <i>P.N. Fedoseyev</i>	3
--	---

INTERSECTIONAL SYMPOSIUM: NEW PATTERNS OF EXPLANATION IN SCIENCE

The Rediscovery of Time, <i>I. Prigogine</i>	29
--	----

INTERSECTIONAL SYMPOSIUM: SCIENCE AND ETHICS

Ethics and Science, <i>E. Agazzi</i>	49
Is There Anything We Should not Want to Know?, <i>P. Gärdenfors</i>	63
The Ethics of Science as a Form of the Cognition of Science, <i>B.G. Yudin</i>	79

1. FOUNDATIONS OF MATHEMATICAL REASONING

Non-monotonic Reasoning by Axiomatic Extensions, <i>G. Jäger</i>	93
Inexact and Inductive Reasoning, <i>J. Paris and A. Vencovská</i>	111
Problems of Admissibility and Substitution, Logical Equations and Restricted Theories of Free Algebras, <i>V.V. Rybakov</i>	121

2. MODEL THEORY

On the Existence of End Extensions of Models of Bounded Induction, <i>A. Wilkie and J. Paris</i>	143
Towards the Structural Stability Theory, <i>B.I. Zilber</i>	163

3. FOUNDATIONS OF COMPUTING AND RECURSION THEORY

Automorphisms of the Lattice of Recursively Enumerable Sets and Hyperhypersimple Sets, <i>E. Herrmann</i>	179
Degrees of Functions with No Fixed Points, <i>C.G. Jockusch, Jr.</i> . .	191

4. SET THEORY

Free Sets for Commutative Families of Functions, <i>U. Abraham</i> . .	205
Polarized Partition Relations and Almost-Disjoint Functions, <i>J.E. Baumgartner</i>	213
A Dilworth Decomposition Theorem for λ -Suslin Quasi-Orderings of \mathbb{R} , <i>M. Foreman</i>	223

5. GENERAL LOGIC

Logic and Pragmatic Truth, <i>N.C.A. da Costa</i>	247
The Justification of Negation as Failure, <i>K. Fine</i>	263
First-Order Spacetime Geometry, <i>R. Goldblatt</i>	303

6. GENERAL METHODOLOGY OF SCIENCE

Strong and Weak Methods, <i>V. Filkorn</i>	319
Impact of Global Modelling on Modern Methodology of Science, <i>J.M. Gvishiani</i>	333
Conceptual Change and the Progress of Science, <i>D. Pearce</i>	351
Scientific Method and the Objectivity of Epistemic Value Judgments, <i>K. Shrader-Frechette</i>	373

7. FOUNDATIONS OF PROBABILITY AND STATISTICAL INFERENCE

The Interface Between Statistics and the Philosophy of Science, <i>I.J. Good</i>	393
Astronomical Improbability, <i>I.J. Hacking</i>	413
Probability in Dynamical Systems, <i>J. von Plato</i>	427

8. FOUNDATIONS OF PHYSICAL SCIENCES

An Axiomatic Basis as a Desired Form of a Physical Theory, <i>G. Ludwig</i>	447
On Learning from the Mistakes of Positivists, <i>G. Nerlich</i>	459

9. FOUNDATIONS OF BIOLOGICAL SCIENCES

Evolution — Matter of Fact or Metaphysical Idea?, <i>R. Löther</i>	481
Evolutionary Altruism and Psychological Egoism, <i>E. Sober</i>	495

10. FOUNDATIONS OF PSYCHOLOGY AND COGNITIVE SCIENCES

Vision and Mind, <i>V.D. Glezer</i>	517
---	-----

11. FOUNDATIONS OF SOCIAL SCIENCES

Rationality and Social Norms, <i>J. Elster</i>	531
On the Nature of a Social Order, <i>I. Pörn</i>	553

12. FOUNDATIONS OF LINGUISTICS

Informational Independence as a Semantical Phenomenon, <i>J. Hintikka and G. Sandu</i>	571
How Natural is Natural Language?, <i>J. Koster</i>	591

**13. HISTORY OF LOGIC, METHODOLOGY AND
PHILOSOPHY OF SCIENCE**

Leibniz and the Philosophical Analysis of Science, <i>F. Duchesneau</i>	609
The Logical Ideas of N.A. Vasiliev and Modern Logic, <i>V.A. Smirnov</i>	625
Phenomenalism, Relativity and Atoms: Rehabilitating Ernst Mach's Philosophy of Science, <i>G. Wolters</i>	641
Contributed Papers	661
Index of Names	693

PRESIDENTIAL ADDRESS

DANA S. SCOTT

*Department of Computer Science, Carnegie-Mellon University,
Pittsburgh, PA 15213, U.S.A.*

Professor Frolov, Distinguished Guests,
Esteemed Colleagues, Ladies and Gentlemen:

It gives me great pleasure to greet you as President on the opening of the 8th Congress of Logic, Methodology, and Philosophy of Science.

There are many people who deserve thanks for doing all the work that such a congress takes. First and foremost, I want especially to thank the Chairman of the Soviet Organizing Committee, Professor Ivan T. Frolov, and his many colleagues and co-workers on the Committee for undertaking the organization of this Congress and for carrying out the work so successfully. Next I must thank the Chairman of the Program Committee, Professor Jens-Erik Fenstad, and the other members of this Committee for the care and interest they have taken in organizing the sections and choosing invited speakers. Also, I wish to give very warm thanks to both the Secretary of the Executive Committee, Professor Risto Hilpinen, and the Treasurer, Professor Helmut Pfeiffer, for all the complex correspondence and details of administration which they have executed so responsibly. You can hear their full reports at the General Assembly, which will take place on Tuesday at 15:00h in this hall. I wish additionally to thank all the National Committees, who have also had very much labor in obtaining and disbursing funds for the speakers and contributors to be able to attend the Congress, and of course I thank the speakers and participants for making the effort to come here.

This Congress is perhaps the most significant activity of the Division of Logic, Methodology and Philosophy of Science. Its organization requires three full years or more, and the work does not stop after the Congress: the members of the Organization Committee, the Program Committee and the Executive Committee have to arrange the publication of the

Proceedings and of other Congress publications, to hand over material and reports to the next committees, and to keep the contacts alive with the national groups and various agencies and organizations. For example, I would like to report that, during the past four-year period, the Division has continued to grow: after the present General Assembly it will number 37 Ordinary Members and 3 International organizations. The Executive Committee is particularly happy to announce the new membership of the Committee of the China Society for the Dialectics of Nature and the China Association for Mathematical Logic in the Division. We also now have four member Organizations in South America (Brazil, Chile, Peru and Venezuela), and the logicians and philosophers of science of at least one more country in South America are endeavoring to create a Committee which could join the Division through a national scientific organization. Inquiries have recently been received from two countries in Africa, and the Committee hopes that the work of the Division can be extended on that continent too. But, again, let me say this is clearly a large amount of work, and the question arises: Why do it? There are two main answers.

In the first place, let us consider the topic of this congress: *Logic Methodology and Philosophy of Science*. How are we to understand these words? May I suggest that they should be grouped together? Do we not want to study the *Logic-Methodology-and-Philosophy of Science*? In *Logic* we find the tools of reasoning, but they must be adapted for various sciences. Logic can be an abstract discipline, but it loses much without its applications. *Methodology* concerns the principles of the organization of knowledge, but each science has special demands on organization. In our lifetimes we have ourselves seen fantastic changes in systematization in physics, cosmology, geology, evolution, biology, linguistics—to name but a few—and also in mathematics. The work is far from finished, and we will live to see many more changes. The very existence of change requires that Methodology be studied. Of course, some will class this study as philosophy, but *Philosophy* for us has wider concerns, in particular, *criticism* and *evaluation*, that are as important as Logic and Methodology. We need to have a balance among all these aspects of the investigation of the development of Science.

Concerning the point just made, I have to remark further on the nature of our organization. We are the Division of Logic, Methodology and Philosophy of Science of the International Union of History and Philosophy of Science. The other division concerns History of Science. We have certain formal connections but not nearly enough contact.

Smaller meetings have been arranged, but the schedules of our International Congresses are different. I hope there will be some discussion at the General Assembly about ways of having greater collaboration and intellectual exchange. It is not possible to work on Logic, Methodology and Philosophy of Science without a sense for the History of Science. Nor is it possible to avoid contacts with working scientists. Our Program Committees have done well to bring interesting speakers from many subjects, but we must try more to have smaller joint meetings with scientists of many kinds. Departments at universities and academies are too often *compartments* with doors and windows closed. An organization such as ours can do much to open up communication on all the problems of common concern.

The second main reason for having international meetings is that they are *international*. I remember the 1st Congress at Stanford in 1960 very vividly. At that time there had not been so many congresses, and it was a great opportunity to meet many interesting people. We have now had large numbers of congresses in many subjects since then, but in our area there is a special reason for international gatherings. Logic, Methodology and Philosophy of Science cannot be separated from consideration of the Ethics of Science, and in this realm there is a great need for international understanding. The Program Committee and Organizing Committee have kept this in mind, but so should we as participants.

All over the world intellectuals have a great many privileges, in particular, freedom of thought and freedom of travel. True, at one time or the other these freedoms have been restricted, but in the long run the privileges exist. My point is that *privilege engenders responsibility*. The enormous and very fast development of Science has put the world in great peril. Over this century we have been in the grips of terrible wars: political, economic, religious, and racial. The climate of war has caused too great a growth of the economy of armament. We have to remember what is called "defense" can too easily change to aggression. The only hope for our world is greater communication and understanding between peoples. We of the Division of Logic, Methodology and Philosophy of Science, both in personal contacts and through our scholarship, can contribute in essential ways to this communication. *We must!*

In this spirit of international freedom of thought,
I declare this congress open!

This Page Intentionally Left Blank

Inaugural Address

This Page Intentionally Left Blank

PHILOSOPHY, SCIENCE AND MAN

PYOTR FEDOSEYEV

Academy of Sciences of the USSR, Moscow, USSR

Science as a concentrated and theoretically systematised experience of humanity, as an organic product of its culture, has in the 20th century been exerting an ever greater influence on the development of man and society. Hence, philosophical reflection has to comprehend the development of knowledge in the context of that interaction.

In our epoch, the philosophy of science is increasingly oriented to analysing the growing role of scientific and technical progress in the life of man and society, inasmuch as the dynamic character of modern social development is largely a result of the scientific and technological revolution.

Philosophical knowledge, regardless of the subjects it deals with, has always turned, or returned, to man. Although some philosophical theories tackled being in general, they all have ultimately concentrated on the problems of the being of man. If philosophy was founded on the spiritual principle, right up to the absolute spirit, according to Hegel, it absolutised and projected human consciousness, man's spiritual creation on to the world. When the material principle is taken as the basis, an analysis leads, in some way or other, to an understanding of the objective conditions of man's existence and development.

Both the sciences of nature and those of society are essentially aimed at studying the world around man, and the natural or social conditions of his being and development. Geology or astronomy seem to study our planet Earth or the Universe without any connections with man, but the results and data produced by these sciences are of vital significance for man. It is precisely this circumstance that has given a powerful impetus to their progress.

In Marxist philosophy the social and humanitarian significance of science is regarded as its basic characteristic. Science is considered as

both the product and the driving force of the universal historical process of mankind's development.¹

Outstanding natural scientists have always regarded solutions of the problems of man to be the ultimate aim of science. W. Heisenberg, for instance, acknowledged that "natural science always presupposes the presence of man. . . . The object of study by natural science is not nature as such, nature by itself, but nature as an object of human problems. . .".²

The humanitarian orientation of the natural sciences was comprehensively substantiated and resolutely emphasised by the eminent Soviet chemist Academician N.N. Semenov. He wrote: "The ultimate goal of the natural sciences is to create the best possible condition for human existence. Science fulfils this mission with the help of the human desire to cognise the surrounding world, the mysteries of the structure of matter and of the laws governing its motion. Everything that man has achieved in the material sphere, from obtaining fire to the use of atomic energy, is due primarily to that wonderful desire to cognise the world around us."³

The sciences of society and the various aspects of its life, its structures and functions exerting everyday influence on men are connected with the interests of man all the more closely. Since these sciences directly bear on the interests of men, social groups and classes, their development proceeds in quite a contradictory way; quite often they have not discovered the truth so much as distorted and concealed it. Philosophy has also suffered that impact. Unfortunately, and for those same reasons, the achievements of the natural sciences were to a great extent used against humanity, particularly in the destructive wars of the 20th century.

However, genuinely scientific knowledge about both nature and society is created by man and ultimately discloses the conditions of his existence. This is why the question of the relationship between science, man and society is the crucial one for the elaboration of philosophical, methodological and logical problems of scientific knowledge.

It should be noted that the understanding of the role of the human factor is increasing all over the world, and among the philosophers of science as well. It is no wonder that the positivist concepts of scientific knowledge, which were reduced to purely logical analysis, are being replaced by new approaches which go beyond the abstract methodological constructions and take into account the fact that science is man's creation.

¹ MARX, K., and ENGELS, F., *Works*, Russian 2nd edn. (Moscow), vol. 1, p. 292; vol. 26, part I, pp. 355, 399-400.

² HEISENBERG, W., 1956, *Das Naturbild der heutigen Physik* (Hamburg), pp. 12, 18.

³ SEMENOV, N.N., 1981, *Science and Society*, Russian edn. (Moscow), p. 170.

In this connection the science of science treats in a new way the role of philosophy as the school of human thinking and comprehension of the world.

Discussions have long been held about the correlation of philosophy and natural science, about the borderline between them, about philosophy's contribution to the development of science, and the role of natural science in the development of philosophy. Some scientists maintain that philosophy and natural science differ by their objects of investigation, and that philosophy should not study natural objects. Others believe that the boundary between philosophy and natural science lies in the fact that the objects and problems under study are examined by philosophy and natural science from different angles. Without going deep into this discussion, I would like to emphasise the main point: it is not the opposition, but the interaction of philosophy and science that is a condition of their fruitful development. Genuine scientists were well aware that natural science could not get along without philosophy. The great Russian natural scientist V.I. Vernadsky, the founder of biogeochemistry, wrote the following about the role of philosophy in the development of knowledge as far back as 1902: "It seems to me that these are aspects of one and the same process—the aspects quite inevitable and inalienable. . . . If one of them ceased to exist, the growth of the other would also stop. . .".⁴ It is precisely such an orientation to the need for an alliance between philosophers and natural scientists that was expressed and thoroughly substantiated by Lenin.⁵

In its development, philosophy relies not only on the richest experience of its own but also on the achievements of the natural and social sciences. For its part, it can play an important role, first, in the generalisation and interpretation of the latest scientific achievements; secondly, in the integration of knowledge and the creation of a general scientific picture of the world; thirdly, in the perfection and development of the methodology and logic of scientific knowledge; fourthly, in an analysis of the socio-ethical problems of science in relation to man, society and nature. In short, philosophy has great cognitive, methodological and conceptual significance.

Let us examine, albeit briefly, these functions of philosophy implemented in its interaction with science.

The philosophical analysis, generalisation and interpretation of the

⁴ VERNADSKY, V.I., 1981, *Selected Works on the History of Science*, Russian edn. (Moscow), p. 7.

⁵ LENIN, V.I., *Complete Works*, Russian edn. (Moscow), vols. 18, 45.

latest concrete scientific data not only establish their connections to and differences from the knowledge accumulated earlier, but also lay the methodological foundations of the formation of a new system of scientific views.⁶ These generalisations can be made at a level of philosophical principles and categories as such, and also at a level of general scientific concepts and conclusions. Being intermediate between the philosophical and concrete scientific levels, this level of general scientific knowledge has a host of unused possibilities for its elaboration by both specialists in concrete branches of science and philosophers. Works by Professor Ilya Prigogine, a Nobel Prize winner, vividly demonstrate the great conceptual significance of generalisations made at the level of general scientific knowledge by a natural scientist. Thus, as a result of the disclosure and scientific generalisation of a broad range of unbalanced states and unstable structures, and of the elaboration of the principle of self-organisation, our notions about the processes current in the world and the law-governed patterns have changed radically.⁷ Philosophers could more actively utilise the corresponding opportunities given by scientific generalisations for a productive synthesis of such philosophical conclusions which anticipate the promising development trends of concrete scientific knowledge and serve as its theoretical reference points and value standards.

Philosophical conclusions drawn from scientific discoveries concretise and enrich the general scientific picture of the world, thus exerting a direct influence on both the principles and categories of knowledge, and the conceptual orientations of people.

The philosophical synthesis of the picture of the world produced by integrating modern knowledge of nature and society makes it possible to unite the latest scientific data about the Universe and the structural organisation and developmental processes of matter; the general concept of the global historical process; the comprehension of the global problems of our epoch as a specific manifestation of the interdependence of nature and humanity, as well as the interdependence of different societies,

⁶ Academician B.M. Kedrov provided an apt interpretation of the role of philosophy in scientific cognition: "Philosophy, of course, cannot offer a chemist or a biologist a concrete method for studying precisely chemical or biological phenomena. It has also dealt with the ideal general method of thinking in accomplishing any scientific tasks which can bring the researcher to the correct result, so that he could reveal a previously unknown truth. In other words, it shows what the general road to discovering truth must be." KEDROV, B.M., 1972, *Philosophy as a General Science in Its Relationship with Particular Sciences*, Russian edn. (Moscow), p. 418.

⁷ PRIGOGINE, I., 1980, *From Being to Becoming* (San Francisco).

countries and peoples at their present development stage; the humanistic understanding of man as the subject and an end in itself of the historical process. I should note that the general scientific picture of the world is recognised by and elaborated in Soviet philosophical literature not only as a major conceptual achievement of the theory of knowledge, but also as a foundation of the scientific outlook.⁸

As for theoretico-cognitive problems, the principal question of the philosophy, methodology and logic of science has for quite a long time been, and remains, the question as to whether they can analyse the genesis of new scientific knowledge and the prerequisites and possible ways for discoveries, or are they able to check, explain and substantiate the truth and reliability of existing scientific conclusions and hypotheses?

Logical positivism has set as its main task the study of the problems of the verification of scientific knowledge, and denied the possibility of analysing the ways in which new scientific ideas come into being.

At present, another view has gained currency, according to which methodology could produce something more.⁹ New thinking is forcing its way into the science of science. Calls are heard ever more frequently for working out the methodology of scientific work and discovery.

In our view, the new thinking in the science of science boils down to the development of dialectical thinking oriented not only to determining the criteria of truth, but also to a creative scientific quest, to solutions to new problems of science and to practical activities in the interests of man.

Such an approach is incompatible with narrow empiricism and pragmatism, for it requires a thorough theoretical analysis and substantiation. At the same time dialectical thinking rejects the natural-philosophical claims which force on natural science *a priori* systems and hypotheses, separated from reality, or reject without proof some scientific concepts.¹⁰

⁸ *Scientific Picture of the World. The Logical and Epistemological Aspect*, Russian edn. (Kiev, 1982); *Scientific Picture of the World as a Component of Contemporary World Outlook*, Russian edn. (Obninsk, 1983); *Scientific Picture of the World. General Cultural and Interscientific Functioning*, Russian edn. (Sverdlovsk, 1985).

⁹ A relevant discussion was held during the Nevada Conference (USA). *Case Studies* (Dordrecht, 1980), vol. 1, *Scientific Discovery*; vol. 2, *Logic and Rationality*.

¹⁰ All-Union seminars on philosophical aspects of natural science (1958) and on philosophical aspects of higher nervous activity and psychology (1962) had a great significance in the practical overcoming of the incompetence of the nature-philosophical interference in scientific progress. *Philosophical Aspects of Natural Science*, Russian edn. (Moscow, 1959); *Philosophical Aspects of Higher Nervous Activity and Psychology*, Russian edn. (Moscow, 1963).

Take for instance cosmology. Some astronomers and astrophysicists believe that cosmic bodies are formed from scattered, diffusive matter through its compression. Others insist that the evolutionary processes are developing in the opposite direction: from a dense and overdense state to a less compressed one. It is clear that the question of the nature of the matter from which the observable cosmic systems have been formed and of the mechanisms of their formation, is a question of natural science, an astronomical and astrophysical question. It should be solved on the basis of a comparison with the data obtained by observation. It is quite possible that in the course of time one of the competing concepts of evolution will triumph or their synthesis will take place in one form or other. The problem discussed, however, has an essential philosophical aspect. Philosophers want to know what is the general direction of the processes of cosmic evolution — is it unilinear, going always and in all cases only in one direction, or is there a dialectical interaction of processes going in opposite directions? Why can we not assume that the cosmic formations in some conditions evolved from more dense states to less dense ones (as was the case with our metagalaxy at the early stage of its formation), and that in other conditions it evolved in the direction of a greater density of matter? Philosophy does not solve those problems but it draws one's attention to them, emphasising that it is the study of the interaction of opposite processes in different conditions that will help understand the appearance in the outer space of superdense and scattered states.

It is from this very point of view that we regard the problem of the finiteness and infinity of the Universe and its cognition. The finitist conceptions of scientific knowledge (current in Western Europe, the USA and the USSR) which rely on scientific data, particularly on the models of the Universe provided by relativist cosmology, insist that physics and the theoretical reconstructions based on it have proved the finiteness of the Universe and allegedly reached the "absolute" limits of knowledge of the laws of nature.¹¹ However, the experience of both history and science suggests that nature is not restricted by the limits of our present-day knowledge of it, that it is endless in its diversity and development. The knowledge of it is inexhaustible and will be continually enriched, deepened and renewed. In solving this particular problem philosophy teaches us the wisdom that the metric models of a finite Universe reveal only one aspect of our knowledge of it. Mathematicians themselves have

¹¹ PIETSCHMANN, H., 1983, *Das Ende des naturwissenschaftlichen Zeitalters* (Frankfurt am Main).

long since proved that the measurable continuum, within which the said models are build, consists, even in a strictly formal sense, of a countless multitude of points.

Thus, dialectics warns us against the absolutisation of any concepts, the dogmatisation of the level of knowledge achieved, and reveals new prospects and new tasks for research.

The point is not to create a certain set of logical instruments for the deductive generation of new theories, but to practice a conceptual-cognitive orientation instilling confidence in the infinite possibilities of human knowledge, going from limited relative truths to a fuller and more profound knowledge.

The principal theoretical and practical conclusion stemming from that premise is that dialectical thinking is the all-round development of the culture of creative thinking.

Science began to free itself from the fetters of dogmatism, authoritarianism and monologue in the sphere of human spirit from the time of Galilei. It would be wrong, however, to assume that the process of casting off these shackles has ended. In both modern science and philosophy the manifestations of dogmatism and authoritarianism are still felt to this day. This is why the philosophical comprehension of the changes taking place in modern natural science should become a school for a new, "dialectical", creative thinking.

Philosophy as a conceptual-methodological nucleus of new thinking should answer the challenge of our time. It is faced with the task of renovating itself and becoming capable of comprehending the processes and problems of the modern world as an entity, including man and society. Apparently, not a single philosophical school is satisfied now with the role of its concepts in the development of science and culture, society and humankind. As for us, we view very critically the state of affairs in the philosophical science. In this inner discontent lies an effective stimulus for the further development and renovation of philosophical knowledge, and its growing role in the life of man, society and entire humanity.

In this context we attach great importance to the study of the history of philosophy and logic, science and technology, which accumulates the experience of human thinking and knowledge.

The development of the culture of creative thinking in accordance with the requirements of a dynamic socio-economic advancement and scientific and technical progress underlies the reform of secondary and higher education now being implemented in the USSR. The restructuring of

scientific research, the democratisation of life and of the activities of scientific institutions, societies and associations also proceed in the same direction. Of course, the culture of creative thinking can develop successfully on the basis of definite theoretical requisites.

We consider the concept of the unity of the world and, accordingly, the unity of scientific knowledge, which has substantiated the objective law of the integration of the methods and achievements of various sciences, an essential gain of dialectico-materialist philosophy. The universal principle of development, including the emergence of new qualitative states and the principle of the interconnection of phenomena and processes, is the determining factor of this concept.

This trend toward the unity of knowledge was noted by Karl Marx; he voiced the idea about the formation in the future of a uniform science embracing both nature and society and oriented towards man.¹² Soviet philosophers are actively elaborating this range of questions.

The unity of science is now openly recognised by the Western science of science, although it sees its source not so much in the objective world as in the human requirement for the creation of a single paradigm of explanation for the whole reality.

The integration of knowledge, according to the dialectical conception, is objectively predetermined by the material unity of the world, which is manifested in the infinite multiformity of natural and social processes, their interrelationships and contradictions.

The substantiation of the unity of the world has been a theoretical prerequisite of the recognition and analysis of the general laws of the development of nature, society and human thinking. The principal cognitive significance of the discovery of the unity of certain basic features of developmental processes both in the objective world and in human thinking lies in that it explains the possibility of an adequate reflection in human consciousness of objects and phenomena in the surrounding world, that is, the possibility to comprehend the objective scientific truth. The connection between human thinking and natural processes was thoroughly analysed by Engels and Lenin.¹³ Our comprehension of science as a developing objective knowledge about the world is based on

¹² MARX, K. and ENGELS, F., *Works*, vol. 2, p. 166; vol. 3, p. 16; vol. 12, pp. 4, 727–728; vol. 42, p. 142; vol. 26, part I, pp. 355, 399.

¹³ MARX, K. and ENGELS, F., *Works*, vol. 20, pp. 10–14, 22–25, 35–36, 87–94, 366–372, 516–519, 526–527, 528–583; LENIN, V.I., *Complete Works*, vol. 18, pp. 5–6, 34–63, 97–146, 244–251; vol. 29, pp. 98–99, 160–165, 169–171, 176–178.

that analysis. The denial of the objective truth, in our view, weakens and undermines the positions of science and opposes the accelerating scientific progress. This is why we attach great importance to the recognition of the objective truth as a condition of man's adequate and ever more thorough knowledge about the macroworld and the microworld, as the moral stand of the scientist.

It should be noted that Western philosophers ever more often criticise antirealism and acknowledge that science deals with the objective reality. Professor R. Harre, for instance, writes that the antirealism which infiltrates society as antisience is not only erroneous but also morally inconsistent.¹⁴ He accepts the view of J. Aronson who believes that theories are reflections of independent phenomena which are directly observed by the researcher, and of the objects which can be studied using only the necessary instruments and deductions by analogy, because they are linked with general ontology.¹⁵

However, some philosophers of science who have discussed the scientific status of the various trends of irrationalism and relativism, have stated that rationality is a historical phenomenon having no real grounds.¹⁶ Scientific rationality is of course a relative and historically changing phenomenon like our entire scientific knowledge which develops from relative truths to a more exact and complete knowledge, but the fact is that even a relative truth reflects objective reality, and relative knowledge contains objective truth. It is with the objective truth that we connect realism and rationalism in scientific knowledge, and reject, on that basis, all and sundry forms of antirealism, irrationalism, relativism and "methodological anarchism".

The growing requirement for the integration of knowledge conditioned by the unity of the world calls for the broad development of comprehensive interdisciplinary research. The interaction of natural, technical and social sciences and humanities becomes closer and more intensive, and the general trend of scientific knowledge as a whole is to tackle the problems of man.

Research oriented to problems that change the traditional trends and forms of the activity of scientists has become increasingly important in

¹⁴ HARRE, R., 1986, *Varieties of Realism. A Rationale for the Natural Sciences* (Oxford, New York), p. 6.

¹⁵ ARONSON, J., 1984, *A Realist Philosophy of Science* (London).

¹⁶ There even exists a committee of scientific studies of paranormal claims. FRAZIER, K., 1984, *From psychic and ESP beliefs to UFO's and ancient quacks*, Highlights of CSICOP's First International Conference (Buffalo), vol. 8, No. 3.

science in the latter quarter of the 20th century. For instance, a trend toward a synthesis of specialisation and universalisation of scientific work has been well pronounced in modern biology and the technical sciences. Genetic engineering, the study of the biospheric processes and the elaboration and design of modern technologies are characterised not only by a synthesis of theoretical knowledge and the methods of various disciplines, but also by a fusion of theoretical and experimental investigations, which requires from specialists a combination of formerly separate forms of research activity. New theoretical knowledge obtained in studying an object is often transformed during the process of research into an outline of an experiment and even a design of further development.

The integration processes make us view the traditional problems of the philosophy and methodology of science in a new way. For a long time methodological problems have been investigated predominantly with an orientation to the requirements of fundamental sciences. Today, the range of methodological investigations has become much broader. Along with the methodology of fundamental sciences, whose development poses for philosophers and methodologists quite a few new tasks which are connected with the revolution taking place in these sciences, specific problems arise in the sphere of applied sciences and scientific and technical development. The latter problems may acquire primary importance for the philosophy and methodology of modern science, if one takes into account the significance of interdisciplinary research, as well as the modern synthesis of scientific achievements with questions of design.

Current engineering activity has posed a number of major methodological problems related to our world view. First of all there is the interpretation of the essence and significance of technological activity. Natural science deals with objects created in a "natural way", that is in the course of the development of nature itself. Technical systems are products of man's purposeful constructive activity, of his hands and intelligence. The perfection of the technical means which we have created is the criterion of the depth of our penetration into the mysteries of nature.

Science is the tomorrow of production. Branches of natural science, as they develop, become branches of production. It is in this way that atomic engineering, radiotechnical and electronic industries, laser technique, microbiological technology, and the production of synthetic materials have appeared. This interaction between natural science, technology and production is a characteristic feature of the current scientific and technological revolution.

The prior development of the fundamental branches of sciences of nature provides the necessary pre-conditions for scientific and technical progress, for new machinery and technologies. On the other hand, the development of the scientific-technical disciplines, engineering and design affect natural science, mathematics and even social science. Computerisation facilitates such an interaction. Thus, the nature of the present-day scientific work cannot be explained without an analysis of the foundations of the constructive, technical and technological activity.

Under the conditions of the fusion of fundamental and applied investigations and the intensive exchange of ideas between various disciplines, the activity aimed at forming a synthesis of the ideas about the object under study and the methods of its cognition, begins to play a special role. Questions inevitably arise about the interrelationships of the subjects of various sciences from which the original ideas are borrowed, about the connection between the subject and the method, about the limits and possibilities of the use of the methods tested in one field of knowledge in a completely new field. The same holds for the ever greater drawing together of the logic, methodology and philosophy of science.

The establishment of the unity of science does not exclude its differentiation and growing division into individual disciplines. Moreover, this process is just as legitimate and natural as the integration of knowledge. It stems from the qualitative multiformity of the world, indivisible in its essence. However, one cannot but see that the continuous division of the cognitive process into isolated spheres, disregarding their common ties, leads to a disintegration of knowledge. This is true of both the sciences about nature and the sciences about society and man. The division and isolation of scientific disciplines hampers scientific progress. Social science suffers most from such negative consequences.

The differentiation, and in some aspects the disintegration, of social knowledge is based on the increasing diversity of the human world: the societies and institutions, social groups and organisations comprising it, culture and mass consciousness, the diversity of human personalities, ever more often referred to as unique personalities. The growing diversity is a progressive phenomenon, the source of the further development of the human race as a whole and of the emergence within it of new, deeper interconnections and interdependencies. That process, however, was often interpreted by social scientists not in its contradictory integrity, but one-sidedly, from the point of view of some of its isolated aspects. The whole was ignored and there appeared a tendency towards the differentiation of the social sciences which was apt to lead to disintegration. This

was observable not only in the relations between various fields of social knowledge (such as the economic, legal, historical and other sciences) but also in the relations between special areas within a number of fields which were formerly integrated enough; in particular, in the case of specialities studying general questions in the given field and the concrete specialities in the same field of knowledge. Political economy, for instance, became alienated from concrete economic disciplines, and world history from the history of individual countries and epochs. In sociology, the long desired unification of the general sociological theory, special theories and empirical studies has not been realised, but not because the need for their integration has not been comprehended. On the contrary, such a comprehension existed and even great efforts were undertaken to realise the necessary integration. Various meetings were held on the methodological aspects of the social sciences, but all this failed to bring about any notable results.

There were of course some objective obstacles; among them the character of the organisational structure of the functioning of the social sciences, which quickly and effectively secured differentiation of knowledge and put hardly passable barriers on the road of integration. In addition, the cold war impeded contacts among social scientists on a global scale and restricted their activity to national and regional limits.

Philosophy also has its share of responsibility for that because it failed to reveal the foundations of the growing diversity of social structures and social sciences, their hidden unified, and at the same time, contradictory essence. Philosophy in its development followed specialised scientific knowledge, registered the new differentiations it revealed but did not discover the ever deeper foundations of that differentiation. Neither did it ensure a synthesis of differentiated knowledge.

Consequently, the tendency toward differentiation prevailed in philosophy itself: the various trends and schools in modern philosophy spearheaded their intellectual potentials against each other rather than tried to comprehend the essential foundations of their differences and confrontations. Particular philosophical elaborations got a higher status than generalisations of a fundamental philosophical nature within the philosophical schools themselves.

To enhance its integrating role with respect to concrete sciences philosophy will have to synthesise anew the very foundations of philosophical knowledge. That presupposes a new interpretation of major achievements of philosophical thought throughout its entire history, particularly in the 19th and 20th centuries, a more profound understand-

ing of the fundamental philosophical principles of man's attitude to the world such as the principle of the material unity of the world, the principle of its development and the principle of contradiction as a source of development. Some philosophical categories also require a new interpretation, such as the subjective and the objective, necessity and chance, possibility and reality, and also practice, activity, creativity. New philosophical principles and categories may be introduced and some general scientific notions may be granted a philosophical status.

We have in mind not the further elaboration and certain renovation of each philosophical principle and category separately but the establishment of such an interconnection between all of them, which would express the methodology and logic of present-day scientific knowledge in the most adequate way. We have in mind the development of integrative synthetic trends in science, and the establishment of dialectical connections between different scientific disciplines, each of which is oriented to the problem of man in its certain aspects.

Philosophical-methodological knowledge is organically included in scientific research and provides conditions for a successful solution of special tasks. Take, for example, informatics. Its range of problems includes the study of the mechanisms of the coding, translation, storage and use of the most diverse types of human knowledge. Modern informatics not only studies the processes of the coding and processing of information, but also analyses its cognitive value. The specific problems of informatics connected with an analysis of the information concepts of knowledge, their transmission and creative reconstruction are closely related to the philosophical problems of the social nature of knowledge, the interrelationships of individual and social consciousness, and the forms and methods of the preservation of knowledge in culture. Philosophical elaboration of these epistemological and methodological problems contributes to a synthesis of the natural scientific and socio-humanitarian aspects of modern informatics.

A typical feature of science today is the inclusion of its humanitarian component into modern natural and technical sciences which is revealed, for instance, in a radical change in the character of engineering activity, which is gradually being transformed from the traditional field of projecting and creating new technical means into projecting and creating the integral complexes of human activity. It is no longer the machine but the man-machine system, with its emphasis on the human factor, that is becoming an object of present-day engineering. A complex of ergonomic, ecological, social and psychological parameters should be taken into

account in projecting such a system. New technologies which are the product of present-day engineering activity are intended for coordinating natural and technical knowledge proper with the knowledge of a socio-psychological and humanitarian character.

The projecting and introduction of such technologies should account the consequences of their impact on man and his natural and social environment. Therefore, their elaboration is in essence becoming the research and projecting of "man-machine system-environment" complexes. It is not surprising that many large-scale technological elaborations are becoming at the same time scientific and technological programmes covering research not only into technological, but also ecological, social and psychological problems. This has resulted in large-scale research programmes uniting the efforts of computer and physico-chemical sciences and socio-humanitarian disciplines (psychology, sociology, ethics, law, linguistics, etc.).

Similar large-scale programmes uniting the cognitive possibilities of various disciplines are now emerging and being worked out in connection with solving global problems: ecological, demographic and medical problems; problems of space development; and the creation of new sources of food, energy and raw-materials.

New knowledge obtained in the process of developing such programmes exerts active reverse influence on fundamental scientific disciplines, thus opening new opportunities for their progress. The elaboration of large-scale scientific and technological projects demands ever more acutely the synthesis and systematisation of philosophic and methodological knowledge which would create an integral picture of research activity and coordinated knowledge of various disciplines (physical, chemical, ecological and the like) with social values.

As comprehensive research is becoming a main road for developing present-day science, it evokes a number of problems pertaining to the philosophical interpretation of the integrative processes of present-day scientific knowledge. These problems relate to the process of shifting the boundaries between certain disciplines, and to the interrelation, dialogue and mutual penetration of natural and socio-humanistic knowledge. Until recently studies of the questions of logic, methodology and philosophy of knowledge were oriented towards natural science as a single model of scientific knowledge in general. Today an analysis of the philosophical and methodological problems of socio-humanistic knowledge is becoming imperative, because of their growing intrinsic importance and because their solution is necessary for understanding the development of scientific knowledge as a whole.

Adopting the achievements of all the fields of knowledge, the theory of cognition itself is becoming ever more comprehensive in its character. Considering this process, Lenin noted in this connection that the theory and logic of cognition should be derived from the whole life of nature and intellectual development, and shaped on the basis of the history of knowledge as a whole: the history of philosophy, the history of certain sciences, and the development of language; with special consideration of psychology, the physiology of sense organs, and the mental development of children and animals.¹⁷ Today much of this, but far from all, has been realised in this sense in the development of the theory of cognition.

Turning to an analysis of the ideological significance of philosophy and science, we should emphasise in this regard the importance of the general scientific picture of the world which they produce as a single whole, including its infinite manifestations and the interrelations between natural and social systems and cultures.

I would like to draw attention to the fact that today science is playing an ever more important philosophical role, as it discusses those questions which in the recent past were in the domain of the discussions of philosophers only. Problems of the place and role of man and his consciousness in the world, the functioning and development of scientific activity, the structure and mechanisms of cognitive process, the ethics of scientific research — all this has become today a subject-matter of specific sciences. In these conditions the interaction of philosophy as world outlook and philosophy as the methodology of scientific knowledge not only increases, but also acquires new social dimensions, since it is the philosophical conception of science that determines the attitude of society to science, the methods of social, cultural and moral control of its development, and the ways of using scientific achievements. Today the question of the social responsibility of those developing philosophical conceptions of science is as pressing as the question of the social responsibility of the scientists.

The questions of world outlook and methodology have become so closely entangled that they share their problems. The problems concerning the ontological character of scientific theories, the problems of truth, reason and rationality, and those concerning the interrelation between knowledge and activity, are essentially philosophical problems and simultaneously most important in working out the questions of the logic and methodology of science. There are surely many special methodological problems in science which do not reach directly the philosophical level.

¹⁷ LENIN, V.I., *Complete Works*, vol. 29, pp. 80, 314.

However, they are related in one way or another to pivotal questions of the philosophy of science. Philosophical reflections on science mould the self-consciousness of science, and promote a better understanding of its possibilities and prospects, the mechanisms and moving forces of the growth of scientific knowledge, the character of its relations to other forms of social consciousness, mode of life and culture.

The organic inseparability of science and its philosophical and methodological foundations has been realised for a long time, but it has become especially evident at the present stage of science's development. Today science experiences, more frequently and to greater degree than in previous times, the periods of radical change of fundamental concepts and notions which are usually called scientific revolutions. If in classical science of the 19th century they could be considered, to some degree, extraordinary situations, now they have become normal conditions of research when we consider science as a whole, as a system of interacting disciplines. As science discovers new objects and phenomena it has to remake its foundations for ensuring the examination of new objects. A cursory glance at the history of natural science of the 20th century will suffice to show a succession of scientific revolutions embracing its fields one after another. The emergence and development of quantum-relativist physics, the revolution in cosmology which is due to the discovery of the instability of the Universe, the development of genetics and cybernetics, the information explosion in science, and the intensification of the integration of natural, social and technical sciences—all these phenomena are revolutions determining the history of 20th century scientific progress.

Modern natural science is living through a period of intensive revolutionary transformation. This is true with respect to the entire range of natural sciences: the physics of elementary particles and solid state physics, astronomy and chemistry, biology and geology, geography and ecology. These transformations are due to the essential changes in the methods of research and first of all due to the intensive use of computers in cognitive process.¹⁸ However, the main point is that they touch upon the very foundations of science. To characterise integrally the changes in modern science, one can say that they are to a considerable extent linked

¹⁸ Computerisation produced a sort of world-outlook and ethical problems and even talks about "computer ethics". WHEELER, J.A., 1982, *The computer and universe*, Int. J. Theor. Phys. 21 (7), pp. 557–572; FEYNMAN, R.P., 1982, *Simulating physics with computers*, Int. J. Theor. Phys. 21 (7), pp. 487–488; 1985, *Metaphilosophy* (Oxford), 16, p. 14.

with the establishment of a new system of scientific concepts which are expressed by using such notions as non-linearity, self-organisation, complexity, irregularity, spontaneity, multi-level, purposefulness, globality, and so on.

It should be first emphasised that the philosophical problems of modern natural science are those of development, particularly the problems of understanding its regularities. Prigogine writes: "A profound conceptual reorganisation of science is going on. Revealed everywhere are processes of evolution and diversification (emergence of diversity), instability. We are well aware that we live in a pluralistic world where we meet phenomena both deterministic and stochastic, reversible and irreversible."¹⁹ It should be also added that the philosophical problems of modern natural science are complex, interdisciplinary and ultimately global problems.

The theory of development is enriched by adopting the ideas of our compatriot V.I. Vernadsky, the eminent scientist and thinker who united the geological history of the Earth with the history of everything living on it, and on that basis developed the idea of the noosphere, according to which intelligent human activity becomes the main determining factor in the development of material processes on the Earth.

In the restructuring of the foundations of science during such revolutions, philosophical-methodological ideas are necessary for the critical interpretation of the traditional ideas on the subject and of the methods of science, and a prerequisite for working out new long-term strategies of research.

Current studies in the logic, methodology and philosophy of science have reached the very essential conclusion that to understand how real science is functioning and developing, it is not sufficient to study the structure of its language and theories, the interrelations between various components of scientific knowledge, and the effects of certain methodological results as something that exists independently. It is necessary to analyse the human parameters as well.

The issue here is that scientific knowledge is produced, disseminated and developed in certain human, i.e. social and cultural forms. The same pertains to the perception of scientific knowledge and to intrascientific communication. Scientific programmes originate and develop in the

¹⁹ PRIGOGINE, I., 1987, *Prospects of the studies of complexity*, in: *Systems Research: Methodological Aspects. A Yearbook*, Russian edn. (Moscow).

framework of definite philosophical pictures of the world which bear the imprint of a specific historical, social and cultural environment.

Today the problem of society's impact on the development of science has become the subject-matter of various disciplines: economics, sociology, social psychology and the history of science. It is not without reason that the so-called cognitive sociology of science has emerged and developed considerably and that some of its representatives make attempts at solving many philosophical and methodological issues.²⁰ Recently this problem has attracted the attention of specialists in the logic, methodology and philosophy of science. It is not fortuitous. Science is the most dynamic cultural force created by man. In the middle of this century, which engendered the scientific and technological revolution signifying the confluence of scientific and technical progress into a single flow, science has turned into not only the most dynamic but also a very powerful force of society. Today, science is able to meet vital social requirements and produce an answer to them, demanding, in its turn, the proper social possibilities for itself. The construction of an integral image of science as a historical result of integral socio-cultural development created, in its turn, quite a number of complicated problems, which relate first of all to the socio-cultural consequences of the scientific and technological revolution.

Scientism and anti-scientism, the two quite different philosophical positions, give diametrically opposite answers to this question. Scientism believes that scientific knowledge proper always bears a positive cultural value and due to this simple circumstance a steady growth of knowledge is automatically capable of solving all problems and antinomies of culture. Under the present conditions such an orientation means practically completely uncontrolled development of science and its concentration on the purely cognitive tasks. Anti-scientists, on the contrary, insist on a principal confrontation between science and culture and explain all the troubles of society by the development of science and technology. They see the salvation of culture in limiting their expansion. Anti-scientism comes close here to scientism. For both of them, science does not exist as a human force born by man and for man, which develops in social and cultural forms. It is regarded by both as something alienated from man and in a certain sense opposing him, thus forming in essence a dehumanised image of science.

²⁰ For relevant polemics see: R. HARRE, *op. cit.*

We Marxists proceed from the ideas of the objectivity, rationality and truth of scientific knowledge. It is namely the Marxist philosophy that always underlines the necessity of comprehending scientific knowledge as a part of the social, cultural and historical context, of a certain system of forms of human vital activity.

It is worth noting that the synthesis of the ideas about the objectivity of the truth of scientific knowledge and its social, cultural and historical conditionality is not a simple matter. How to preserve the understanding of the fact that scientific knowledge characterises a reality independent of our consciousness, and simultaneously theoretically interpret the cultural and historical changeability of its forms and content, scientific methods and the very ideas about the ideals and norms of scientific character?

Some researchers believe that in solving the problems arising here, an analysis of the questions of the methodology and philosophy of science should be separated from the social and cultural context. Starting from a real and important fact of transforming science into a direct productive force, the fact of blurring the stable, old-world boundaries between fundamental and applied studies, they explain the whole science solely as a supplier of new technologies. From this point of view scientific knowledge cannot provide the understanding of the profound essential characteristics of reality, but is merely reduced to projecting new types of technological activity (directly or indirectly).

Another interpretation underlines the decisive importance of social conditions for working out a scientific picture of the world and researching specific problems. Scientific theories are considered in this case as purely ideological constructions which depend completely on a concrete socio-cultural situation and have no logical and methodological advantages in relation to magic and mythological conceptions.

Both points of view proceed from relativist ideas, from the opinion that science is incapable of possessing the objective truth. Both positions essentially contrast science with philosophy and do not produce an adequate idea about scientific knowledge.

In this regard I would like to stress that the key to solving these far from simple questions lies in the dialectical interpretation of the interconnection between cognitive and socio-cultural determinants of scientific activity, with due account of the relative independence of science and its social determinants.

The development of science is determined first of all by its inner regularities, by the logic of continuity and innovation in research. At the same time scientific progress depends on how great is society's need for

scientific knowledge, for knowledge of a certain kind and content, on society's readiness to support science directly or indirectly: to finance research, to encourage scientific activity morally and materially, etc. Today this circumstance becomes particularly significant. Today research often requires gigantic material outlays, technical means and large teams of scientists. It is quite clear that modern society is not inclined to look as an outsider at what the scientists are doing, in which direction science is progressing, and what sort of knowledge it produces.

The question concerns the functions of science in the social being of people rather than the direct determination of the results of scientific research by society. It is a question of the role which science as a whole and various aspects of scientific activity play in the multifaceted life of society, and of the demands to which science responds in one way or another, and which therefore determine its development.

In current philosophical studies much more attention is paid to the analysis and role of natural science in society's social and intellectual life. There are many reasons for that. They include the orientation of society's present-day rapid technical and technological advance, environmental protection for the sake of the forthcoming generations, and many other global problems linked with the rational use of the sources of man's existence and development, primarily the overcoming of the atomic challenge to civilisation.

Humanised science, developing as an organic part of the activity of life, plays not only the most important practical and technological role but begins to play an ever more important direct socio-cultural role. I would like to stress here that in today's culture, essential importance is attached to the methods, specific to science, of obtaining universally significant knowledge about the world: intellectual, but non-violent, compelling arguments, criticism, democratism, historical experience in dialogues in search for the truth, etc. This side of the matter is directly related to humanising the social connections of science and to the question of the influence of science on society. This point of view makes it clear that as social and cultural institution, and as a type of public activity, science exerts its ever more pronounced impact on society.

In its essence science is intolerant of otiosity and stagnation, the antipodes of progress. Neither does it accept subjectivism and voluntarism which are incompatible with the recognition of objective truth. From this point of view it becomes evident that science as a social and cultural institution, as a form of consciousness and social activity, has an increasingly great impact on society.

The destiny of the whole civilisation depends on the ability of man to humanise the scientific development. The heated discussions on the questions of the ethical control of scientific research and the social responsibility of scientists are not fortuitous today.²¹

The social functions which science is assuming today present new requirements to the scientific community in ethical and cultural terms. It is not always that science and scientists fulfil those important functions, and this sometimes causes negative assessment of science's humanitarian potential which grows into a general negative assessment of the role of intelligence in general. Felt here are manifestations of the technocratic tendencies, individualism and casteism, the irresponsible attitude to designing and using complex technical devices which greatly harm both people and environment, the indifference to human needs, concerns and destinies. Having become an extremely mighty force, science is capable of destroying humanity if it escapes moral control, and on the other hand, being included into harmonious social and cultural development, it can make a great contribution to the process of civilisation. The latter kind of science is capable of being a forum of wide international social contacts, which make it possible to develop a general view of the most acute present-day problems. This, for instance, is the role of science in predicting the possible destructive consequences of a nuclear war (the "nuclear winter" concept), in working out ways of overcoming the ecological crisis, etc. The most urgent problems of today are preserving peace and civilisation on Earth and protecting nature.

We cannot close our eyes to the fact that science is intensively being entangled in the arms race, militarisation of space and the implementation of the star wars programme. At the same time it is pleasant to note the growing activity of scientists of various specialities in the movement against the threat of nuclear self-destruction of the humanity, and for a lasting peace on Earth. Prominent scientists launched this movement.

Niels Bohr was among the first physicists who realised that the atomic weapon is a challenge to the entire human civilisation, and necessitates a radically new approach to international relations. In his Open Letter to the United Nations of June 12, 1950, he drew attention to the need for an international agreement which would guarantee universal security, and stressed the significance of mutual understanding, mutual trust and stable

²¹ A comprehensive analysis of the socio-ethical problems of science and of the scientists' social responsibility, and also a review of relevant discussions are to be found in: FROLOV, I.T. and YUDIN, B.G., 1986, *Ethics of Science*, Russian edn. (Moscow).

cooperation among nations. The vital necessity of concerted efforts to avert the sinister threat to civilisation opens, in his opinion, an exclusive opportunity to overcome international contradictions. He attached great significance to greater openness in technical progress and in military and political affairs, and emphasised the importance of timely consultations between countries to find the best ways to jointly achieve security. Niels Bohr wrote that the top priority should be given to achieving an open world in which the role of each nation would be determined only by the measure in which it is able to promote general cultural progress and to help other countries with its resources and experience.²²

The voices of outstanding scientists sounded ever louder with every passing year, warning about the deadly consequences of the nuclear rivalry and calling for an immediate intensification of the efforts to avert a nuclear catastrophe. The Russell–Einstein Manifesto which appeared in 1955 on the initiative of the outstanding physicist F. Joliot-Curie, is still relevant. It said in particular: “To preserve life on our planet we, representatives of human beings, should learn to think in a new way and make practical steps excluding wars and the arms race.”²³

Igor Kurchatov, the outstanding scientist, head of Soviet atomic physicists, repeatedly and insistently emphasised the need to prohibit atomic and hydrogen weapons, to develop a broad international cooperation in the peaceful use of atomic energy. In his speech at a session of the USSR Supreme Soviet on March 31, 1958 he said: “We, Soviet scientists, are greatly alarmed by the fact that there is still no international agreement on the unconditional prohibition of atomic and hydrogen weapons. Our scientific community has resolutely come out for outlawing the use of nuclear weapons. Together with the Soviet scientists are foreign scientists of world renown: Niels Bohr of Denmark; Joliot-Curie of France; Pauling of the USA; Heisenberg of FRG; Yukawa of Japan; Powell of Great Britain; and many others.

From this high rostrum we, Soviet scientists, address scientists of the whole world with the appeal to direct and unite efforts to realise as soon as possible controlled thermonuclear reaction and to turn the energy of hydrogen nuclear fusion not into a weapon of destruction but into a powerful life-bearing source of energy bringing welfare and joy to all people on Earth.”²⁴

²² 1950, *Science* 112, 2897, pp. 1–6.

²³ 1960, *Einstein on Peace* (New York), p. 633.

²⁴ KURCHATOV, I.V., 1984, *Selected Works*, Russian edn. (Moscow), vol. 3, p. 198.

The specialists in philosophy, methodology and logic of science together with the progressive public should contribute to the assertion of a new style of philosophical and scientific thinking aimed at establishing the loftiest humane ideals and values and, consequently, at excluding war from the life of society and at developing fruitful international contacts.

The struggle of ideas in philosophy and science will obviously continue in the future. By the way, in accordance with dialectics, such a struggle is an engine of progress. However, the struggle of ideas and their material and practical implementation should not end in military confrontations. Socialism suggests that this method of resolving contradictions should be replaced by the strength of example, the attainment of success in the peaceful competition for scientific, technical and social progress.

I would like to recall that the profound reconstruction of all spheres of public life on democratic principles and wide openness now under way in the Soviet Union and guided by peace and humanism, make focal the human interests and the individual's all-round development. The development of Soviet science is recognised as the most important factor in this revolutionary reconstruction of a great historic dimension.

In conclusion I would like to emphasise that the integrity and humanism of scientific knowledge necessitate, naturally, an active international scientific cooperation. We favour this wide and fruitful scientific cooperation. Historical experience shows that international scientific cooperation, an active exchange of ideas and discoveries, is an important source of progress of the entire humankind and of each nation.

This Page Intentionally Left Blank

**Intersectional Symposium:
New Patterns of Explanation
in Science**

This Page Intentionally Left Blank

THE REDISCOVERY OF TIME

ILYA PRIGOGINE

*Faculté des Sciences (CP 231), Université Libre de Bruxelles, B-1050 Bruxelles, Belgique and
Center for Studies in Statistical Mechanics and Thermodynamics (RLM-7), The University of
Texas at Austin, Austin, TX 78712, USA*

1. Introduction

I would like to dedicate this communication to the memory of my friend Boris Grigorovitch Kuznetsov. His book *Reason and Being* has recently appeared in English translation (KUZNETSOV 1987); once more I was struck by the similarities the approach which was developed by Kuznetsov presents with the one I will present here. The main problem which Kuznetsov singles out is the role of probability and irreversibility in our conception of the Universe. Different approaches to this question are well exemplified by the three following excerpts from Einstein (1916), Kuznetsov (1987), Lucretius (~ -60 BC):

Lucretius:*

*Illud in his quoque te rebus cognoscere avemus,
corpora cum deorsum rectum per inane feruntur
ponderibus propriis, incerto tempore ferme
incertisque locis spatio depellere paulum,
tantum quod nomen mutatum dicere possis.*

Einstein:

The weakness of theory may be summarized by the fact that it does not correspond to the wave theory and that, on the other hand, the time and direction of the elementary processes are determined by chance. Besides, I am convinced by the potentialities of the method I have chosen.

*Translation: *while the first bodies are being carried downwards by their own weight in a straight line through the void, at times quite uncertain and uncertain places, they swerve a little from their course, just so much as you might call a change of motion.*

Kuznetsov:

However, the world-line without certain equivalents of the Epicurean clinamen, without ultrarelativist filling, is not true being, but rather determinate nothing.

These sentences suggest some remarks. Let us first notice the analogy between Einstein and Lucretius. What is emphasized is that the precise time of elementary processes is determined by chance. They both say that without a certain element of stochasticity we would have what Kuznetsov calls in his book "at most a vacuum", but not the world as we know it. In other words, the basic problem is the conflictual situation between the static description proposed by classical physics, based on deterministic and time-reversible laws, and the world as we know it, which for sure includes probability as well as irreversibility as basic elements.

Obviously, the classical view expresses a dualistic structure: the phenomenological level corresponds indeed to irreversible and stochastic laws, while at the fundamental level, classical or quantum, we would have time-reversible, deterministic laws. For the case of quantum mechanics, I refer of course to the description in terms of the Hilbert space. Can we overcome this duality, and attain a more integrated view of physics? I think this is now possible. Of course, the results I will present do not answer all the questions Kuznetsov did ask, but they hopefully represent a step in the direction he had outlined.

The problem which we have to face is a very complex one, and implies a deep conceptual change, which is going on at present. I shall start with the phenomenological, thermodynamical level; next I will consider the changes we have to adopt concerning the languages of classical and quantum mechanics; then we will conclude with some reflexions about the recent evolution of cosmological ideas. Obviously, the range of these problems is enormous, and therefore I should apologize for the somewhat superficial character of the remarks I will develop here (more details are to be found in the original papers (BROUT *et al.* 1978, GÉHÉNIU and PRIGOGINE 1986, GUNZIG *et al.* 1987, GUNZIG and NARDONE 1987, PRIGOGINE and PETROSKY, 1988a, b, c)).

This being said, it is not by chance that I have here to treat such a wide range of physical phenomena; in fact, I would like to communicate you my impression that a more unified physics, based on concepts such as stochasticity and irreversibility, is now in the range of our possibilities. However, many things will obviously have to be added or changed in the description I shall present.

2. Phenomenological description

It may be appropriate to start here with the second law of thermodynamics. The evolution of entropy in an open system is split into exchanges with the environment of the system (entropy flow $d_e S$, which may be positive, negative or zero) and an internal entropy production $d_i S$ (which corresponds to irreversible processes, and is always positive or zero) (see Fig. 1).

Let us stress the importance of irreversible processes (PRIGOGINE 1980, PRIGONINE and STENGERS 1983). As a matter of fact, we know that all of chemistry, all of biology is made of irreversible processes. In addition, I should like to emphasize here that we know now that irreversibility is not only related to destruction of structures, to disorder: entropy production involves both order and disorder. As a single example, let us consider a well-known physical effect: thermodiffusion. Take a closed system with two components, hydrogen and nitrogen. At uniform temperature, there is also a uniform distribution of hydrogen and nitrogen. If one imposes a thermal constraint on the system, introducing a gradient of temperature, one observes a gradient of concentration. We see that entropy production has indeed a double effect: it is associated to a heat flow, producing disorder, but it is also associated to anti-diffusion, and anti-diffusion means order, as it produces a partial separation of hydrogen and nitrogen. This double effect can be observed in many situations (see Fig. 2).

We are really here in front of a new paradigm. The tradition associated order to equilibrium and disorder to non-equilibrium, as exemplified by the contrast between crystal and turbulence; but we have now reached an opposite point of view, in which the creation of order is associated to non-equilibrium, while disorder may be associated to equilibrium structures, even to crystals if we include the description in terms of normal

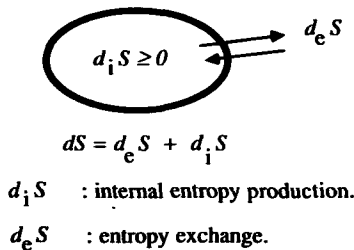
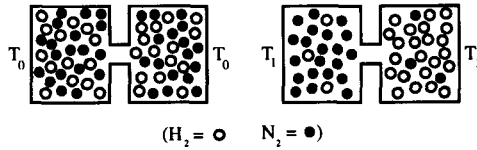


Fig. 1. Thermodynamical relations for open systems, which present both an internal entropy production ($d_i S$) and an exchange of entropy with the environment ($d_e S$).



$$\frac{d_i S}{d t} = \begin{array}{l} \text{Thermal flow} \\ \geq 0 \end{array} + \begin{array}{l} \text{Antidiffusion} \\ \leq 0 \end{array} \geq 0$$

$$\frac{d_i S}{d T} = \text{"order"} + \text{"disorder"}$$

Fig. 2. Thermodiffusion effect: under thermal constraint ($T_1 \neq T_2$), the distribution of H_2 and N_2 becomes inhomogeneous.

modes. This creation of non-equilibrium structures is well known here in Moscow, as the Russian school has played an important role in their exploration in many situations such as chemical oscillations or hydrodynamical instabilities.

Let us consider the Rayleigh–Bénard instability. If I have chosen this example, it is because it has paved the way to some recent numerical experiments. I refer to the work done by MARESCAL and KESTEMONT (1987, 1987) on molecular dynamics of non-equilibrium systems. Usually, molecular dynamics deals only with equilibrium systems; but the recent years have witnessed a development of simulations in non-equilibrium fluids, due to the huge increase of computing power offered by ‘super-computers’.

The work of Mareschal and Kestemont deals with the Rayleigh–Bénard instability we have just mentioned. The system consists of an assembly of 5400 hard disks enclosed in a rectangle; vertical sides are reflecting boundaries, whereas horizontal sides are thermal reservoirs. An external force (similar to gravitation) acts downward on the fluid’s particles, and the temperature of the bottom’s reservoir is set higher than the top’s reservoir’s one. The behaviour of this model fluid is integrated over time on a computer. Figures 3b and 3c refer to the same system after about 10^6 collisions. They differ by the thermal gradient, which is larger in Fig. 3c than in Fig. 3b. Figure 3b corresponds to constraints below the hydrodynamical macroscopic instability. Still, coherent patterns are already present. However, they fluctuate violently in time (see Fig. 3).

It is quite remarkable that a relatively small system exhibits a behaviour which can be understood in terms of macroscopic hydrodynamics. However, the results of the simulation go beyond hydrodynamics proper,

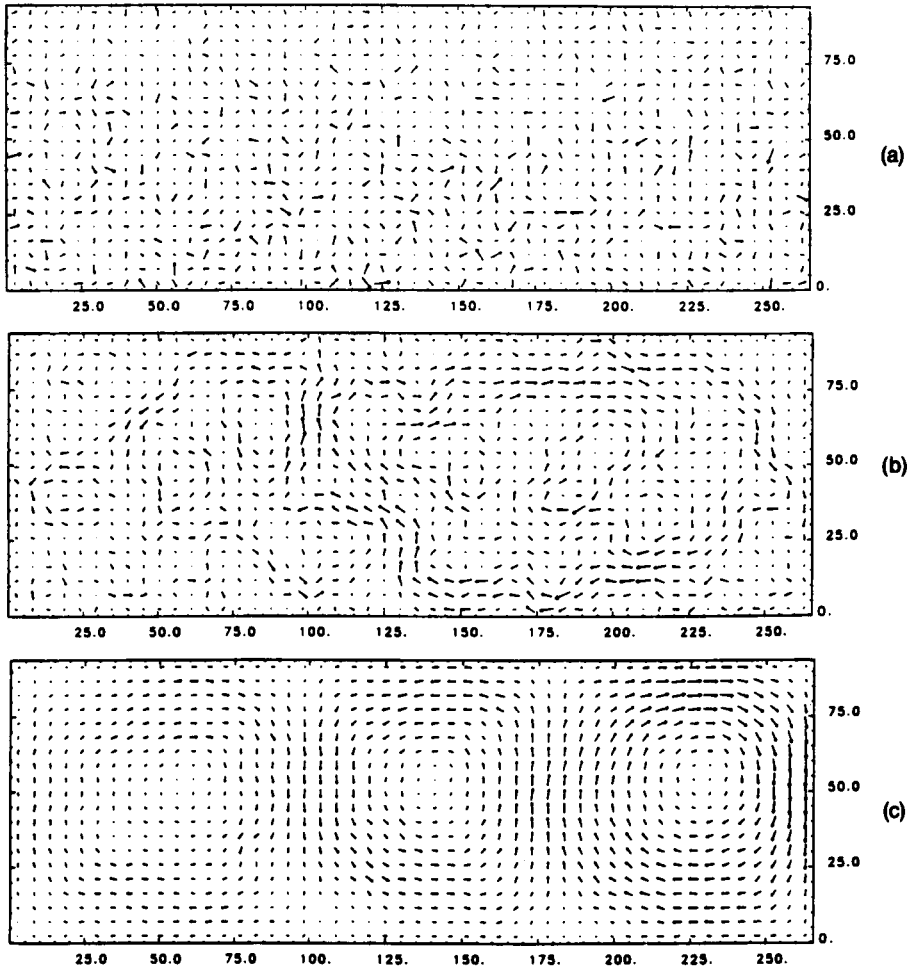


Fig. 3. Finite two-dimensional system of 5400 hard disks under the influence of a thermal gradient and an external force (gravitation). The figures show a typical velocity field: each arrow of the graph is an average over a cell of the velocities of the particles that belong to the cell (typically 5 per cell). Picture (a) presents initial (random) mean velocity densities. Picture (b) presents the same mean densities after a few thousand collisions; vortices corresponding to regular flows of particles can now be seen. Picture (c) presents the same system under different external constraints; while whirlpools seem indeed to be present, there are no stable ordered structures.

as they show how the macroscopic instability is prepared by fluctuations. Long-range correlations have also been shown recently to play a role in chemical systems, as shown by work under progress by M. Mareschal and A. Amellal. They lead to spectacular effects; for example, the fluctuation

of a region which is small in respect to the correlation length follows different laws than fluctuations in large regions. This example shows how non-equilibrium may generate long-range correlations. The fact that non-equilibrium may play a constructive role is now established beyond any reasonable doubt. I will now turn to the microscopic description of irreversible processes, in the conceptual frame of classical and quantum mechanics.

3. Conceptual changes at the microscopic level

Here the situation has dramatically evolved over the last three decades. In the book I wrote with Isabelle Stengers, I stressed the fact that irreversibility comes from dynamical instability. Today, this idea is generally accepted. As an example, I should like to quote a recent paper by J. LIGHTHILL (1986): *I have to speak on behalf of the broad global fraternity of practitioners of mechanics. We collectively wish to apologize for having misled the general educated public by spreading ideas about the determinism of systems satisfying Newton's laws of motion that, after 1960, were to be proved incorrect.* It is quite unusual to see a scientific community presenting apologies for a mistake which had lasted for over three centuries.

The most extreme cases of dynamical instability is given by the Kolmogorov flows; we may measure the degree of their instability through the "Lyapunov exponents", according to which the distance between two trajectories increases exponentially; this exponent gives us the "temporal horizon" of the system. This implies that dynamics deals now with systems presenting an intrinsic stochasticity (see Fig. 4).

$$\delta x_t = \delta x_0 e^{\lambda t}$$

$\lambda > 0$ (Lyapunov exponent)

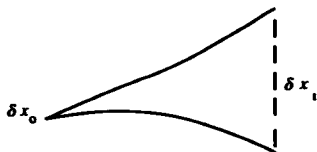
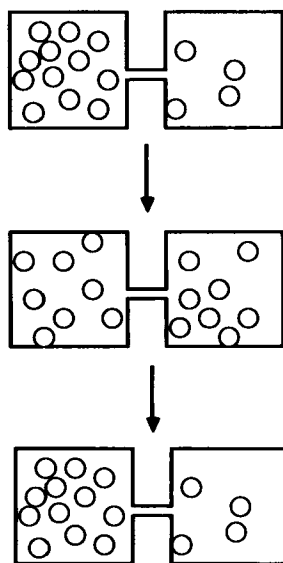


Fig. 4. Positive Lyapunov exponents measure the divergence of trajectories starting from nearly identical initial conditions.

I want also to emphasize briefly the fact that the kinetic description, which goes back to Boltzmann, is very closely related to dynamical instability. We cannot go into details here, but I want to mention the important role played by resonance; resonance is really—we know it since Poincaré's theorem of 1892—what prevents a dynamical system from being integrable. It is because of resonance, as manifest in the so-called problem of the small denominators, that some classical systems proved to be non-integrable. As everybody knows, this is the starting point of the Kolmogorov–Arnold–Moser theory. Now, a fundamental concept of kinetic theory is the collision operator, and this operator is directly related to resonances. This question has been recently studied in collaboration with my colleague T. PETROSKY [see the papers by PRIGOGINE and PETROSKY (1988a, b, c)]. So, kinetic theory is very close to Kolmogorov flows and we may formulate here the same remarks about the existence of a “time horizon”.

We may now see why the classical statements about the illusory character of irreversible description are no more true. One of the main arguments used was of course Poincaré's “recurrence” (see Fig. 5).

Poincaré's recurrence Theorem



Irreversibility only "apparent"

Fig. 5. The Poincaré's recurrence Theorem presents the return of any integrable system to its former state after a sufficiently long time.

According to this view, every dynamical system should be periodical in the long term; we should recall here Smoluchowski's conclusion as presented by Weyl: "If we continued our observation for an immeasurably long time, all processes would appear to be reversible" (quoted by WEYL 1949). For a sufficiently long time, every system would be quasi-periodic. The difference between irreversibility and reversibility would thus be only one of the scale of observation. As stated by Chandrasekhar: "we may conclude with Smoluchowski that a process appears irreversible (or reversible) according as whether the initial state is characterized by a long (or short) average time of recurrence compared to the times during which the system is under observation" (CHANDRASEKHAR 1949). In other words, there would be nothing like true irreversibility in nature. But we now understand how to avoid such a conclusion. As Poincaré's recurrence time is generally immense in respect to the Lyapunov temporal horizon, Poincaré's recurrence Theorem is not applicable for highly unstable systems. Indeed, beyond Lyapunov's temporal horizon, the concept of a trajectory is destroyed and we have to use a more suitable description. Other paradoxes, such as the Loschmidt's paradox, can be discussed in the same way.

However, I would prefer to make some remarks about the existence of an arrow of time even for large dynamical systems at equilibrium. Let us consider more closely the collisional mechanism in a collection of hard spheres. In such a system, we can make a distinction between "pre-collisional" correlations and "post-collisional" correlations. A few years ago, BELLEMANS and ORBAN (1967) presented numerical results using molecular dynamics, leading to the evolution of Boltzmann's \mathcal{H} quantity. They started from an uncorrelated ensemble of hard disks, and have shown, in agreement to earlier calculations, that the \mathcal{H} quantity was decreasing monotonously in time. They then inverted the velocities, after a given number of collisions. In agreement to Loschmidt's paradox, the \mathcal{H} quantity then presents a period of increase, comes back to about the initial value, and decreases again, see Fig. 6. How do we understand this result? In the "direct" evolution, collisions randomize the velocities, and create post-collisional correlations, which are then destroyed by subsequent collisions. In the inverse evolution, obtained by inverting velocities, we have pre-collisional correlations leading to a decrease of the randomness of velocities. However, this corresponds only to a transient situation. For long time, we come back to a mechanism increasing the randomness of velocities. What is then the status of Loschmidt's paradox? If we could proceed with velocity inversions for arbitrary times, there

could be no privileged direction of time. However, this is precisely what the existence of Lyapunov time horizon prevents us to do. We can only invert velocities for times shorter than or comparable to the Lyapunov time. This simulation shows that the statements by Chandrasekhar and Smoluchowski we just quoted are incorrect. Whatever the duration of the observation, there is a direction of time, leading to equilibrium in our future.

An interesting point following the new simulations by Kestemont and Mareschal is that there is an arrow of time even at equilibrium. They show that while before collision, molecules at equilibrium are uncorrelated, collisions do generate correlations. Collisions create correlations, which then die out. This is not so astonishing. Indeed, collisional mechanisms remain the same. After all, colliding molecules do not know that the system to which they belong is in equilibrium or not.

We see therefore an arrow of time even in an equilibrium system. This is quite unexpected, because, of course, it seems to imply some violation of micro-reversibility. The issue is a privileged direction of time, corresponding to the sequence [collisions \mapsto correlations], and not to the sequence [correlations \mapsto collisions]. This is not trivial, as, let us say, for a 2-bodies system in a finite container, this would not be the case (see Fig. 6).

Therefore, we have an arrow of time, but which has no macroscopic consequences at equilibrium. The “violations” remain on the microscopic

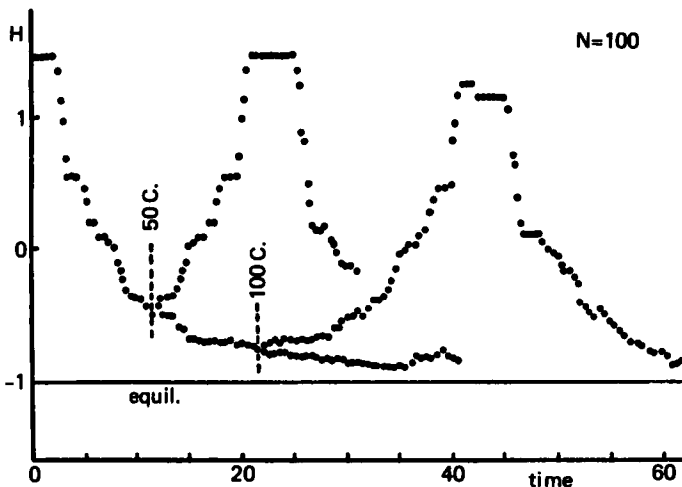


Fig. 6. Evolution of \mathcal{H} over time for a system of 100 hard disks. Velocities are inverted after 50 collisions (open circles) and after 100 collisions (solid circles).

level, precisely because the system is at equilibrium. However, if we impose external constraints on the system, leading to non-equilibrium conditions, we would allow the arrow of time to appear at the macroscopic level. We are thus in a situation quite the opposite from the one which was described by classical theory: the preparation of the system does not introduce the arrow of time; but it allows the arrow of time to have macroscopic effects. The arrow of time existed before; but it was hidden, neutralized by the very specific constraints needed to maintain the system at equilibrium.

4. New perspectives in quantum theory

I would like to give here some indications about some recent directions of research in quantum theory, from the point of view of irreversible processes. In its present state, quantum theory presents a curiously dual structure. On the one side, the wave function evolves in a deterministic fashion in the Hilbert space; this evolution is in addition time-reversible.

$$i \frac{\partial \Psi}{\partial t} = H_{op} \Psi \quad (4.1)$$

On the other hand, irreversibility and stochasticity are introduced by the measurement process, which leads from probability amplitudes in Hilbert space to probabilities proper.

$$\Psi = c_1 u_1 + c_2 u_2 \quad (\text{measurement} \rightarrow |c_1|^2 \text{ and } |c_2|^2) \quad (4.2)$$

Quantum mechanics leads to the paradoxical conclusion that irreversibility would only have a meaning because of human measurement, a situation reminiscent of the status of irreversibility in classical mechanics before the discovery of highly unstable systems. This is one of the quantum mechanics paradoxes. This dual interpretation of quantum mechanics is reasonable, when the process described by the Schrödinger equation is "in itself" a reversible process. For example, if we consider an idealized system like a harmonic oscillator formed by a neutral particle, then indeed the only way to introduce irreversibility in such a system where motion is reversible would be through measurement. Is this the only case? This would mean there are no irreversible events in nature independent from our measurement. How then to avoid the paradox clearly stated by A. Rae: *The "measurement problem" . . . arises from the idea that quantum systems possess properties only when these are meas-*

ured, although there is apparently nothing outside quantum physics to make the measurement (RAE 1986). If irreversibility “enters” through measurement, how to meet situations which appear as intrinsically irreversible, as radioactivity, quantum jumps or the death of Schrödinger’s cat?

It is interesting to notice that in the old Bohr–Sommerfeld–Einstein theory there was room for “intrinsic irreversibility”. Indeed, in the old theory, we had on one side the Bohr–Sommerfeld conditions for the energy levels, given in terms of action variables J characteristic of the theory of integrable systems,

$$J = \oint p dq \quad (4.3)$$

and on the other side the transitions between levels described by Einstein in terms of spontaneous and induced emission; Einstein knew well that in this way he introduced the idea of stochasticity and irreversibility on the level of the system as a whole, the atom including the radiation. This view is expressed in the quotation of Einstein which we have presented at the beginning of this paper. Is it possible to introduce intrinsic irreversibility into quantum theory? Certainly, a deep change would be necessary. This was already predicted in a sense by Eddington when he wrote (EDDINGTON 1928):

The whole interpretation is very obscure, but it seems to depend on whether you are considering the probability after you know what happened, or the probability for the purposes of prediction. The $\Psi\Psi^c$ is obtained by introducing two symmetrical systems of waves travelling in opposite directions of time.

The Hilbert space description is certainly closely related to the problem of time reversibility, and if we want to introduce intrinsic irreversibility, we shall have to give up to some extent this description. This changes the whole algebraic formulation of quantum mechanics. These questions have been treated recently in papers which I wrote with my colleague Tomio Petrosky, and here I can only give some general results. The simplest way to discuss this change seems to be to start here with the quantum mechanical uncertainty relations. Everybody knows the usual quantum mechanical uncertainty relations, which express that we cannot measure simultaneously momentum and coordinate with infinite precision, and as you know, this comes from the fact that momentum and coordinate are related to non-commuting operators. The usual uncertainty relation states that

$$\Delta p \Delta q \geq \hbar \quad (4.4)$$

Its interpretation is standard: it expresses that the operators associated to momentum p , and to coordinate x are non-commuting. Let us next consider the uncertainty relation

$$\Delta E \Delta t \geq \hbar \quad (4.5)$$

Its interpretation is somewhat subtler, as there is no operator corresponding to time. Still, in many situations, there is no real difficulty to give a meaning to this relation. For example, if we take a wave packet, its lifetime will be related to its spectral representation, through this uncertainty relation. We may then consider the case of an unstable state, characterized by a lifetime τ . As well known, we have

$$\Delta E \geq \frac{\hbar}{2\tau} \quad (4.6)$$

However, this uncertainty relation has a quite different meaning. Indeed, the lifetime has a well-defined value for a given quantum state and a given experimental device. Therefore, this uncertainty relation limits the measurement of a single quantum observable, in this case the energy of the unstable quantum state. We may write this relation in the form

$$\overline{E^2} - (\overline{E})^2 \geq \left(\frac{\hbar}{2\tau}\right)^2. \quad (4.7)$$

It expresses a dispersion in the values of the energy. The existence of a finite lifetime leads to a lack of control of the energy of the unstable quantum state, as manifested by the natural line width.

We consider this fact as fundamental, as it suggests that a new form of quantum theory is necessary. Not only do we need non-commuting operators, but we also need operators which would be non-distributive. Indeed, if we represent the observable energy as some operator Λ acting on the hamiltonian AH , we should have

$$\Lambda H^2 \neq (\Lambda H)^2. \quad (4.8)$$

This gives us a hint about the direction in which we have to develop quantum theory to deal with unstable systems. Since the oral presentation

of this conference, progress has been made in the understanding of the basis of the increased stochasticity, which results from matter-field interactions. In short, we start from a bare particle, interacting with the vacuum; the particle then modifies the field in order to be able to transmit its energy through a resonance process. We have here a dynamical self-organization process, occurring over a very short time, typically 10^{-18} s for an atom. This self-organization process leads to a dragged field, which in turn interacts with the unstable particle. In a pictorial way, we could say that the particle is moving in some kind of boundary layer, which it has itself produced. This motion has a statistical character, and can be described by a density matrix (see I. PRIGOGINE and T. PETROSKY 1988).

In other words, the system (matter + light) is a non-integrable one, as can be inferred from Poincaré's celebrated theorem relating irreversibility to resonances. This instability leads to an intrinsic irreversibility, and to a description where the duality between function and measurement is no more present. We may expect the experimental effects due to this intrinsic irreversibility to be small; however, the epistemological frame of quantum mechanics is then drastically altered. In conclusion, let me mention that I agree completely with the general statements made by POPPER (1982) . . . *it may be possible . . . to give an indeterministic reinterpretation of Einstein's deterministic programme, and at the same time an objectivistic and realistic reinterpretation of quantum theory. . . it is likely that the world would be just as indeterministic as it is if there were no observing subject to experiment with it, and to interfere with it.* Let us now discuss the question of irreversibility in the cosmological context (BROUT *et al.* 1975, GÉHÉNIU and PRIGOGINE 1986, GUNZIG *et al.* 1987, GUNZIG and NARDONE 1987).

5. Entropy in the context of cosmology

There is no field of science in which the evolution of ideas has been more turbulent, more unexpected than modern cosmology. As you know, it was in 1917 that Einstein gave the first cosmological model, which was associated to general relativity. This was a grandiose description, but it described a merely geometric, static universe. It was soon established by Friedmann and Lemaitre that the solution Einstein proposed for his own equation was unstable, and that our universe could be originated in a "Big Bang", instead of being static. Later came the measurement of the

red shift, which confirmed the idea of an expanding universe. And more recently, the discovery in 1965 of the black body residual radiation gave us a fossil of an evolutionary universe.

So, over a relatively short time span, we went from a static to an evolutionary universe. Today, there is a generally accepted theory of cosmology, the “standard model”, which is described in many books, and which seems to give a sensible answer to most questions. However, the standard model does not include the very first moments of our universe, the so-called “quantum era”.

According to the standard model, the entropy of the universe would remain constant, while its temperature decreases with the adiabatic expansion of the universe (see Fig. 7). The “creation” of the universe is rejected to a singular point: the so-called “Big Bang”. But if we are to investigate these questions, we have to ask: what is the meaning of the “Big Bang”? Why this initial singularity at all? More generally, how to understand the initial conditions out of which our universe has evolved? These questions seem to be somewhat out of the range of traditional general relativity. A theory which is very popular today is the so-called “inflationary universe”. This theory was indeed successful for some questions left unanswered by the standard model. However, it does not lead to a better understanding of the Big Bang, and it does not lead to a better understanding of the role of the second law of thermodynamics in cosmology. These are the questions which are of interest for us here.

I would like to give here a summary of a scenario which has recently been designed by Gunzig, Géhéniau and myself, and which is based on the work of numerous previous authors among whom I would like to quote Brout, Englert and Nardone. There are of course now many scenarios, and the list becomes longer and longer every month. The reason for devoting some attention to this specific scenario which I will present here is that it leads to interesting predictions, as it does allow us to compute the value of the so-called “specific entropy”, this latter being the

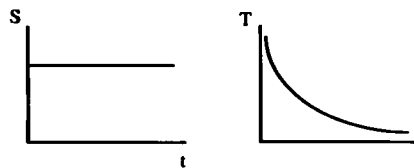


Fig. 7. Standard cosmological model: temperature decreases over time, while entropy is constant.

ratio of the number of photons over the number of baryons. Indeed, the investigation of the residual black body radiation shows that there are about 10^9 photons per baryon. In other terms, the number of baryons is very minute as compared to the number of photons. This of course is a quite striking thermodynamical result, because it shows that most of the entropy of the universe is in the photons. It shows also that it is very likely that the universe has started with a huge "entropy burst".

This is in complete contrast with the traditional interpretation of the cosmological implications of the second law, in terms of which the universe evolves gradually towards a state of maximum entropy (its "thermal death"), starting with a low level of entropy. It seems now that the thermal death is, so to speak, behind us, closely related to the creation of our universe, while the entropy phenomena which are going on today, be they associated to the existence of living organisms, or to the fusion reaction in stars, are just minute if compared to the initial entropy production of our universe.

Let us consider briefly the events which are associated to these views. We start with the so-called "quantum vacuum"; this is not to be considered as a static situation; it is full of fluctuations, full of events such as creation and disparition of particles. Now, a remarkable prediction made by Brout, Englert, Gunzig and Nardone, is that if the mass of the particles which are produced in such fluctuations of the quantum vacuum goes beyond some "critical mass" threshold, of the order of fifty Planck mass, it will provoke a non-linear process resulting in the production of particles and of space-time curvature. A Planck mass corresponds to 10^{-5} g, which means that it is huge compared to an elementary particle. Such a mass can only be a black hole, because the Compton wavelength of this mass is about 10^{20} times smaller than its Schwarzschild radius.

In terms of this model, we may now see how the "initial" vacuum becomes unstable, and shifts into a new stage in which both gravitation and matter are produced in a self-consistent way. In other terms, gravitational energy was instrumental in producing matter. In this sense, we would have a "free lunch", which means that there is neither creation nor destruction of energy, but simply a transfer of negative attractive energy (gravitation) into positive energy (matter). Is this all the story? No, because the matter which is produced in this way has some entropy. As I have just said, the created particles cannot be ordinary elementary particles, which are much smaller. Therefore, these particles can be understood as mini black holes; and from recent literature we know that mini black holes have an entropy, and a life time. Therefore, the novelty

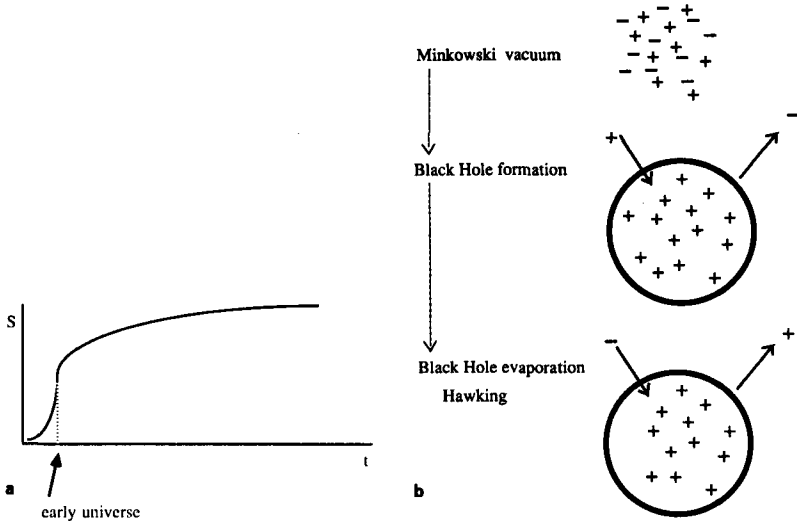


Fig. 8. (a) Evolution of entropy for the model presented here. (b) Transition from Minkowski vacuum to a black hole degenerating into ordinary matter.

of this phenomenon of transfer of gravitational energy to matter is the creation of entropy through the appearance of a mini black holes population. In a sense, this entropy means that these black holes contain, I would say, in a potential fashion, many forms of elementary particles, a large variety of possibilities which will become actual in the subsequent evolution of the universe.

The creation of the mini black holes population would be the first stage of the creation of our universe, followed by the decay of these mini black holes and the apparition of the usual particles, essentially photons and baryons. The lifetime of these mini black holes is of the order of 10^{-37} seconds; the second stage would be the present adiabatic expansion of the universe, while the whole initial production of entropy would have taken place during the decay of the mini black holes (see Fig. 8). As the only constants which are relevant during the initial stages of the universe are the Planck constant h , the velocity of light c and gravitational constant κ , we can predict the present value of the entropy of the universe in terms of these three constants.

According to this description, the creation of the universe is from the start an irreversible process characterized by the creation of entropy. It would be the entropy evolution which could give the characteristic features of the first steps of the universe. Would there not be this entropy evolution, then we could at most speak about reversible fluctuations; and

it is difficult to accept that this universe, which lasts now for about fifteen billion years, is just a reversible fluctuation.

It is also important to notice that the starting point is no more a singularity (as was the case with the "Big Bang" model), but an instability: it is an instability of the vacuum, which leads to a new stage of our universe. Before the universe as we know it, there was another type of phase, the "quantum vacuum", which contains already potentially the universe. I would like to call your attention on the analogy with the role of time at equilibrium in molecular dynamics as we considered it above. There also, there was already time present in an implicit way, as the result of collisions, but with no effect at equilibrium on a macroscopic range. Here, we see that in the pre-universe vacuum, we have a time which is not yet manifest, but was already there in a kind of potential fashion. It becomes manifest when the fluctuations go beyond the threshold established by Brout, Englert, Gunzig and Nardone, and leads then to the creation of the universe. In other words, the universe would be "manifest time", and in this sense the concept of time preexists to the existence of the universe. I am sure that this model which we just considered is not the final one. However, in one way or another, I think that irreversibility plays the fundamental role in the description of nature, starting from the very first stages of our universe.

6. Conclusions

Let us now conclude. I believe that we go further and further away from the classical picture of the physical world, into a direction which is precisely the direction Kuznetsov hoped we could reach, in which we would go away from the idea of the "determinate nothing" which we have quoted at the beginning of this paper. We see everywhere spontaneous activity, irreversibility, non-linearity, stochasticity, fluctuations, instabilities. We can see now some kind of convergence of the world as we experiment it inside us and the world as we see it outside us. This allows us to go beyond the classical duality as expressed by the work of Descartes or Kant, and to reach a type of physics, in which, in agreement with Popper, "The aim is a picture of the world in which there is room for biological phenomena, for human freedom and human reason" (POPPER 1982).

The present evolution of ideas presents a curious coincidence. At the end of this century, we go to a reappraisal of human condition. It is a

striking feature that we also go to a reappraisal of some basic assumptions of physics. To meet this new situation will require new ways of thinking, new observations and mutual tolerance.

Note added in Proof

Since the presentation of this paper, a model of irreversible transition from space-time to matter has been presented (PRIGOGINE I., GÉHÉNIU J., GUNZIG E. and NARDONE P., 1986, Proc. Natl. Acad. Sci. USA 85, pp. 7428–7432). The importance of this approach is that it allows us to avoid specific assumptions which have to be introduced in order to deal with the quantum theoretical approach of relativity.

References

- BROUT, R., ENGLERT, F. and GUNZIG, E., 1978, Ann. Phys. 115, 78.
- CHANDRASEKHAR, S., 1949, "Stochastic problems in physics and astronomy", Rev. Mod. Phys. 15, p. 3.
- EDDINGTON, A., 1928, *Gifford Lectures* (Cambridge, UK), pp. 216 sqq.
- EINSTEIN, A., 1916, Verh. Deutsch. Phys. Ges., 18, p. 318.
- GÉHÉNIU, J. and PRIGOGINE, I., 1986, Found. Phys. 16, 437.
- GUNZIG, E., GÉHÉNIU, J. and PRIGOGINE, I., 1987, Nature 330, 621–624.
- GUNZIG, E. and NARDONE, P., 1987, Fundamentals of Cosmic Physics, 2, 311.
- KUZNETSOV, B., 1972, *Razum i Bytie* (Moscow, Nauka); engl. tr. *Reason and Being* (Reidel, Dordrecht), 1987.
- LUCRETIUS, T.-Carus, ~ -60 BC, *De Rerum Natura* 2, pp. 216–220.
- LIGHTHILL, J., 1986, "The Recently Recognized Failure of Predictability in Newtonian Dynamics", in: J. Mason et al., eds., Predictability in Science and Society, a special issue of the Proceedings of the Royal Society, 407, 35–60, p. 35.
- MARESCHAL, M. and KESTEMONT, E., 1987, Journal of Statistical Physics 48, 1187; Nature 329, 6138, p. 427.
- ORBAN, J. and BELLEMANS, A., 1967, Phys. Lett. 24A, p. 620.
- POPPER, K., 1982, *Quantum Theory and the Schism of Physics, Postscript to the Logic of Scientific Discovery* (Rowman and Littlefield, Totowa, New Jersey) pp. 160 and 177.
- PRIGOGINE, I., 1980, *From Being to Becoming* (Freeman, San Francisco).
- PRIGOGINE, I. and GÉHÉNIU, J., 1987, Proc. Natl. Acad. Sci. USA 83, 6245.
- PRIGOGINE, I. and PETROSKY, T., 1988a, *An Alternative to Quantum Theory*, Physica 147A, p. 461; PRIGOGINE, I. and PETROSKY, T., 1988b, *Poincaré's Theorem and Unitary Transformations for Classical and Quantum Theory*, Physica 147A, p. 439; PRIGOGINE, I. and PETROSKY, T., 1988c, *Intrinsic Irreversibility in Quantum Theory*, Festschrift P. Mazur, Physica 147A, p. 33.
- PRIGOGINE, I., and STENGERS, I., 1983, *Order out of Chaos* (Heinemann, London).
- RAE, A., 1986, *Quantum Physics: Illusion or Reality?* (Cambridge University Press, Cambridge, UK), pp. ix sq.
- WEYL, H., 1949, *Philosophy of Mathematics and Natural Sciences* (Princeton University Press, Princeton, NJ).

Intersectional Symposium: Science and Ethics

This Page Intentionally Left Blank

ETHICS AND SCIENCE

EVANDRO AGAZZI

Univ. of Fribourg, Switzerland and Univ. of Genoa, Italy

The autonomy of science

The modern age — historically understood as the time that followed the twilight of the Middle Ages — may be characterized as the age of the emergence of several “autonomies” in different sectors of the spiritual and practical life of man. Such autonomies had in previous centuries known perhaps only one major example, when Thomas Aquinas clearly advocated the full legitimacy of investigations through “natural” reason (i.e. philosophical investigations) as compared to “supernatural” revelation (that constitutes the basis of theology). In this spirit, Machiavelli vindicated the autonomy of politics, Galilei the autonomy of science, British liberals the autonomy of economics, Kant and the Romantics the autonomy of art, and so on. These vindications originally expressed a particular stress laid upon the *specificity* of the corresponding domains, which entailed the determination of *purely internal* criteria for the fulfilment of their restricted and specific goals. Thus, for example, Machiavelli identified politics with the art of “acquiring, maintaining and expanding the State”, Galilei conceived natural science as the “investigation of the true constitution of the world”, A. Smith presented economics as “an inquiry into the nature and causes of wealth”, Kant, and especially the Romantics, elaborated a new concept of the fine arts as a pure creation of beauty, and the consequence was that criteria were explicitly or implicitly provided for judging when an action is politically wise, a statement scientifically correct, a behaviour economically right, and a creative work artistically valid. Of course it would be wrong to say that proposals for circumscribing and specifying the domain and features of certain disciplines or activities had been lacking before that time. What

was new was rather the fact that the borderlines were now meant to express clear-cut “separations”, rather than simple “distinctions”, and that the consequent “autonomy” of the different fields has quickly turned into a search for a kind of “freedom” or “liberation”.

The step from autonomy to *freedom* may be understood in the sense that the admission of autonomy led to the rejection of any form of tutelage or interference coming from “outside” the single domains. This vindication of freedom was understood in different ways and as having different degrees. In one sense it was conceived as an *independence in the criteria of judgment*, such that, e.g., a decision might be considered as politically sound *in spite* of being economically disadvantageous, a behaviour economically profitable *in spite* of being morally objectionable, a picture artistically beautiful *in spite* of being indecent. This obviously means that, in turn, no consideration of economic or moral criteria (to remain with our examples) could improve the political, economic, or artistic *value* of actions or productions which are negatively judged in terms of their own *internal* criteria. A common way of expressing this position is to say that politics, economics and art are “value-free”, and this is also and especially said concerning science.

A second and much more committed sense is the claim that the said autonomy also entails *independence in action*. In the above examples, this would mean that one is entitled to *perform* a political action in spite of its being economically disadvantageous, an economic action in spite of its being morally objectionable, an artistic work in spite of its being indecent. This means that the politician “as a politician”, the business man “as *homo oeconomicus*”, the artist “as artist” — and we can also add the scientist “as a scientist” — are *legitimated* in acting according to the *pure* criteria of their profession, at least to the extent that they are performing *within* this profession.

A third sense consists of not allowing *controls or limitations* to this performance to be exercised by external agents in the name of the protection or promotion of goals or values of a different nature. It is clear that these different meanings of the “autonomy” are in an order of succession which is not that of an *entailment*, since the acceptance of the first does not imply that of the second, and this does not imply that of the third.

Nowadays, the tendency is clearly manifest to reconsider these points, especially since we are confronted with the outcomes of such a process of “liberation”, which has led to several intuitively unacceptable results: the autonomy of many single domains, if pushed to excess, brings them into

serious conflict with other domains. Thus, the imperative of peace and the respect of fundamental human rights are now advocated as limitations to political action; the needs of protecting the environment, of avoiding technological catastrophes and regulating genetic manipulations are producing a demand for the regulation of science and technology; the promotion of social justice is imposing limitations on the unbridled search for economic profit. Therefore, the delicate problem we now confront is that of effecting a critical revision of the said points, without becoming involved in obscurantism, regressive involution, or negation of the positive aspects which are certainly contained in the claims of autonomy and freedom. It is with this problem that we shall be concerned here.

We shall first try to understand in which sense those domains for which autonomy is vindicated are to be considered "value-free". This cannot mean that they are "devoid of any value", or that those who operate within these domains have nothing in view. Indeed no human action (if it is really "human") is performed without a purpose, i.e. without a goal which is considered worthy of being pursued by the agent. In this sense this goal or aim represents a "value" which inspires the action. Moreover, it follows from the examples considered above that the autonomy of certain domains was actually claimed by explicitly assigning to each of them a well-determined specific aim, and by indicating the criteria for evaluating how particular facts, assertions, actions and products ought to be in order for this aim to be pursued in the most satisfactory manner. This stage corresponds to what we have called above the "independence in the criteria of judgement", and expresses a need for analytic clarification against which it would be hardly possible to raise objections.

This stage does not involve moral problems, since it is only related to action in an indirect and *hypothetical* way, i.e. by suggesting which course of action should be taken *if* the specific goal envisaged were the *only* goal. It by no means *implies* either that this be the unique or the supreme goal of human action in general, or that one should disregard the impact that the fulfilment of this goal might have upon the realization of other human aims or values. Those who accept this "implication" make a transition from the first to the second of the above-mentioned meanings of "autonomy", i.e. to the meaning of "independence in action", and become immediately involved in a specific and highly debatable *ethical claim*. To the extent that the third meaning is related to the second, ethical questions also exist in connection with the third sense.

Considering the ends

Let us now consider science. The above mention of science along with politics, economics and art was made with a purpose, but this purpose was not to put them on an equal footing. Indeed, certain questions of principle concerning the relations of these various domains to ethics, are common to all of them, but each domain also has its own characteristic features. As to science, it is useful to distinguish pure from applied science, not because a clear-cut separation is always possible or recommendable in concrete cases, but because these notions constitute two “ideal types” which should not be confused. Both can be considered as endeavours to provide *knowledge*, but in the case of pure science the goal of this knowledge is (to put it briefly) the discovery of *truth*, in the sense of establishing “how things are”, while in applied science this goal is the realization of some action or practical *result*.

Admitting the specific aim of pure science to be the search for truth, it is clearly immune from moral objections in itself (i.e. it constitutes a perfectly legitimate value). The effort of approaching truth by means of sound and reliable knowledge in different specialized fields—a knowledge which may be characterized by the qualities of objectivity and rigour—has given rise to certain prescriptions concerning the activity of the scientist. They constitute what is usually called “scientific methodology”, but have no ethical meaning at all, being simply *instrumental* to the achievement of the cognitive aim of science, i.e. as means for evaluating whether statements, hypotheses or theories may be credited with the ability of providing sound and objective knowledge. Even those who challenge the rules of methodology, e.g. by advocating some form of methodological anarchism, usually justify their position by the claim of its allegedly being beneficial to the progress of scientific knowledge.

Yet some kind of truly moral requirement seems to be implied by pure science, such as the obligation not to manipulate data, the readiness to accept criticism, recognize one’s errors, credit other people’s priorities, or the devotion to hard work. However, these are not *specific* virtues of science proper, but rather human virtues in a general sense (such as intellectual integrity and self-discipline), which find in scientific practice a privileged opportunity of being exercised. In other words, a scientist who manipulates data in order to credit himself with a fictitious discovery is condemnable not so much as a bad scientist, but simply as someone who has cheated, who has tried to reach a personal advantage (whatever it may be) through dishonest means. This is why scientific “deontology”, in

matters of *this* kind, is not really relevant to the relations between science and ethics, since its rules simply reinforce the fulfilment of the *specific and internal* aim of science, and do not come into conflict with it.

Quite different is the situation of applied science. Here the search for truth is only a secondary end, the primary being some practical realization, and this fact immediately implies the possible existence of ethically relevant issues, depending on the particular *ends* any single enterprise of applied science envisages. The point is sufficiently clear in itself not to deserve extended discussion. To put it briefly: while it is in principle morally acceptable to *know* everything, and there are *no morally prohibited truths*, not everything that can be *done* is acceptable, and there exist *morally prohibited actions*. To deny this would be tantamount to denying the existence of ethics and morality, a position which we must consider alien to a discourse like this, which investigates the relations between ethics and science (admitting by that, that ethics exists no less than science). Hence, *from the point of view of its ends*, pure science does not raise ethical problems (and is always morally acceptable), while applied science raises ethical problems: the problems connected with the goals of the different applications, goals to which applied science is essentially instrumental and, as such, cannot be morally indifferent. It would lead us too far to consider examples, which in any case would deserve very careful scrutiny.

Considering the means

It would be too hasty a conclusion to say as a consequence of the above considerations, that pure science can never be morally objectionable. This was said from the point of view of its ends, but the question of *means* must also be considered, and the general principle that the end does not justify the means, also holds for science. Of course, when saying this we are not thinking of those cases in which a scientist might try to take advantage of some dishonest tools for facilitating his work. We are rather addressing the question whether the acquisition of pure knowledge might sometimes require the use of certain means, the moral legitimacy of which could be questionable. The answer is yes. In fact, at least in the case of the *experimental* sciences, truth cannot be discovered simply by thinking, or by watching, but requires the performance of operations, which implies the *manipulation* of the object that is submitted to investigation. This is not an accidental feature, but the very condition for the

existence of any objective experimental knowledge, which always requires the isolation of certain very specific aspects of reality, through an appropriate creation of artificial conditions of observation and testing. The production of these artificial conditions is what is called here "manipulation" in a quite neutral sense. However, manipulation is *action* and not knowledge, and even when the acquisition of knowledge is its explicit aim, it may well happen that a particular manipulating action not be morally admissible in itself. This fact was not very well perceived when the object of manipulation was nature, since any manipulation of nature seemed to be morally acceptable (nowadays there are quite different views on this point not only regarding the manipulation of animals, but also of inanimate nature). However, it became evident when experimental research on man inevitably implied manipulating man (the paradigmatic case being that of medical research), that moral criteria should guide this very delicate practice, since a very general moral principle prohibits treating a man simply as a tool (quite independently of the more elementary requirement not to harm those who are submitted to the experiment). In fact, the moral reflection concerning experiments on humans has been developed for several decades and has produced the elaboration of certain widely recognized and accepted norms. At present, experiments on human embryos and genetic manipulations for pure research purposes, are widely discussed issues that are interesting to consider here because they show that moral problems may arise in the field of *pure science*, and may imply restrictions of its freedom, in spite of its aims being morally unobjectionable.

It is very easy to recognize that these considerations about the ethical relevance of the means may also be transferred without modification to applied science; the moral acceptability of the goal of a particular applied research does not free us from considering the moral acceptability of the means used in performing this research.

Considering the conditions

Our discussion about means has already called attention on the fact that science (even pure science) is not merely knowledge, since it is necessary to "do something" in order to produce this knowledge, and this fact immediately inscribes science within the domain of action, and not simply of reflection.

Among the factors which are usually involved in moral considerations

regarding human actions, the *conditions* of the action also have special importance. They are similar to the means, but differ from them mainly in that the means are tools for directly reaching the end as a *terminus* of a certain action, while the conditions are something which makes the action itself possible, and thereby serve the end only indirectly. This distinction is useful in order to understand that an action seeking the realization of a morally legitimate goal through the adoption of morally acceptable means still remains open to moral questioning until its conditions have been analyzed.

The most familiar example of this kind of problem, which has been discussed with reference to science in the past few years, is that of the allocation of funds for research. Several questions might emerge in this context. One might concern the provenance of the funds: e.g. would it be morally admissible to accept from a "benefactor" funds which we seriously suspect to be the product of criminal activities, such as narcotic trafficking or kidnapping? Problems also exist in much more normal situations: scientific research is fed throughout the world by great supplies of public money, but public money is always insufficient to fully satisfy all the needs of the community. Hence the money allocated to science is necessarily subtracted from other possible destinations such as, say, hospitals, schools, social security and environmental protection. Since the satisfaction of these needs corresponds to the existence of several aims or values, which it is not only legitimate, but even dutiful to pursue, we easily see that a problem of moral choice inevitably surfaces, a problem whose solution implies determining priorities and also limiting or cutting certain scientific projects. Several other problems, besides that of the allocation of funds, obviously surface when we consider the question of the "conditions" for the existence of pure and applied science; these problems are not treated here.

It would be question-begging to say that in these cases the decision criteria are of a social rather than of an ethical nature, for in any case they should serve to determine what *ought to be done*, and this is the typical feature of any ethical question. To rely upon social motivations for answering the question simply means that, in certain cases, we accept that social values play the role of moral standards (which is by no means incorrect, provided we are aware that there should be other moral standards as well). Let us remark that in the special example considered above, the solution of the moral problem may be easier in the case of applied science than in the case of pure science, for it is usually easier to show how an applied research could "compensate" through its expected

results the sacrifices made by the community, while it might prove more difficult to show the same, in the case of the simple acquisition of knowledge. This shows, among other things, how superficial the idea is that ethical problems are typical of applied science while hardly being of concern to pure science.

Considering the consequences

The last point of this analysis concerns the possible *consequences* of scientific research. It is an obvious moral principle that one is responsible for the consequences of one's actions, and therefore has the duty of trying to foresee them to the extent possible. We mean, of course, the unintended consequences, for those that are intended are simply to be numbered among the goals of the action. This problem has come to the focus of ethical discussions regarding science, owing to the dramatic impact of certain unexpected tragic consequences of technological development, and of the generalized concern about the potentially enormous dangers of an uncontrolled growth of this development. The problem is not new in ethics, however, and has led to the formulation of the so-called "double effect principle" in the tradition. This principle strictly applies to those cases in which the intended end of an action necessarily entails morally unacceptable consequences; but in a more or less stringent way it also applies to those cases in which such consequences are to be expected only with high probability. In such cases the first thing to do is to examine whether it is possible to renounce pursuit of the goal—in order to avoid the unacceptable consequences—and if this is possible, it is also morally dutiful to renounce. Here we have a kind of counterpart of the principle "the end does not justify the means", since it is said that "the end does not justify the consequences"; both statements express a criticism of the thesis that all that matters in ethics is one's intention in performing an act.

However, there are situations in which pursuing the goal has the connotations of a moral obligation. In these cases one has to compare the importance of the two values (the value that is served by the action, and the value that is violated by the consequences of the action), and sacrifice the one that is less important or, to put it differently, to "choose the lesser of two evils". A classical case in which this principle is advocated is that of the "therapeutic abortion", where omitting a certain therapy would mean exposing the life of the mother to a serious risk, while

implementing it would *imply* the death of the foetus: losing the foetus is considered as the “lesser evil” (a situation not to be confused with that of killing the foetus as a *means* of saving the mother).

Situations of the kind envisaged here are not rare in the field of applied science, and in many cases they may be treated not as questions of “all or nothing”, but rather in terms of a balance of “costs and benefits”, which enables one to reduce the risk or the impact of the negative consequences, by reducing the measure of realization of the goal. However, discourse about costs and benefits is possible and correct when the alternatives are homogeneous and allow a common unit of measure, but it is much more difficult or even impossible when we are confronted with a real conflict of values. In these cases the problem appears to be inevitably of an ethical nature.

Even though consideration of the consequences is chiefly a problem for applied science, it is not totally alien to pure science, since not so much the discovery of truth as its *communication* may raise moral questions. Already in everyday life it could be said that “telling the truth” may not always be a moral obligation, due to its consequences. For example, it might be right not to tell a sick man the truth about the severity of his illness, if this would seriously lessen his chances of recovery. Or, a man submitted to torture by the secret police of a dictator, who denies having accomplices (to avoid their being persecuted or killed) is not morally blamed as a liar, but is highly esteemed.

The analogy with scientific truth is not that dramatic, yet it is not unusual that scientific discoveries or theories be communicated to the public in a sensationalistic way, accompanied by superficial and gratuitous interpretations, with possible negative impacts on people’s ways of thinking, and on their appreciation of life and values. This is not always a fault of the media, but sometimes also of more or less distinguished scientists, who indulge in superficial popularization or even in a partisan interpretation or extrapolations of the content of science. At a time when science has attained such high prestige and has such a tremendous influence on the thoughts and feelings of men, the honest and morally scrupulous dissemination of scientific truth has become a major ethical imperative.

The plurality of values

A common denominator of the reflections presented here is the recognition of the existence of a plurality of values, none of which can

pretend to be “absolute” (in the sense of being totally disconnected from the others and worthy of being pursued for its own sake above any other consideration); however no “relativism” is entailed by this claim, since values are not said to *depend* upon the different situations, but simply to *apply* to these situations in different ways. Ethics must start with this awareness, which is simply the projection of the evidence that men are pushed to action by a great variety of motivations, which they consider to be legitimate in themselves, while spontaneously admitting that not everything is permitted in following these motivations, so that *value-judgments* are necessary at every point in order to determine the right course of action. Making one single value absolute (be it pleasure, wealth, power, family, fatherland, friendship, beauty, truth, love, religion) would amount to admitting that in pursuing this value anything is permissible. This would simply mean that the truly moral attitude is suspended in that sphere, since no value-judgments in a proper sense would be adopted for directing the course of action, but simply “efficiency-judgments” concerning the best means to be employed to fulfill the pre-established value. Hence, a subhuman way of acting would characterize this sphere.

The correctness of this statement is not suspended even if one admits a hierarchy of values (e.g. that which is implicitly presented in the ordering of the above list). Indeed we have been accustomed to concede that one acts at a subhuman level when one is oriented exclusively to, for example, the search of pleasure or wealth. This is true, but does not depend essentially on the fact that these are values of lower rank; it depends rather on the fact that these are promoted (consciously or unconsciously) to the position of absolute values. To be persuaded, it suffices to remember how many morally condemnable facts have been or could be the consequence of making the higher-ranking values absolute as well. Hence our conclusion is that science too does not constitute an exception to this general rule. If we limit ourselves to considering science as a system of knowledge (i.e. if we consider *only* its contents), science has no ethical relevance. However, as soon as we consider its also being a human activity, i.e. the activity which aims at producing this knowledge, we must conclude that it cannot help being subject to the general conditions of any human activity, that of being guided by choices inspired by value-judgments, which must take the plurality of values into consideration. From this awareness follows what we said about the evaluation of ends, means and consequences in the case of pure and applied science.

The regulation of science

Some corollaries derive from the considerations developed thus far. The first is that specifically ethical limitations and regulations may concern the practice of scientific research. In fact, as soon as we admit that moral principles must govern human actions, we are obliged to accept that everything is *not* permissible, and that, at the extremities of the interval of what is *permissible*, there is what is *obligatory* on the one side, and what is *prohibited* on the other. However, while moral principles and values are of a very general nature, obligations, permissions and prohibitions concern concrete actions, and must be specified through concrete *norms*. The difficulty with the norms is that they cannot very often be the more or less immediate translation of some general principle, since they must apply to complex situations and actions, which are “complex” because they involve the interference of a number of principles and values.

This consideration has a first elementary consequence, i.e. the fact that criteria, standards or norms elaborated for evaluating the conformity of an action with a given *particular* value cannot be extrapolated to the evaluation of its conformity with a different particular value. In the case of science and ethics, this means that moral criteria could not interfere with the internal judgments concerning what has a *scientific* value, and with the criteria for assessing the validity of scientific results. Symmetrically, moral evaluations have to be based upon ethical criteria of judgment, and are, as such, independent of any interference coming from scientific considerations. This is the correct meaning of the reciprocal “autonomy” discussed earlier. A second consequence is that ethics, owing to its generality (which entitles it to regulate all kinds of human actions), has to view the most satisfactory fulfilment of all the human values that may be involved in a certain action. This means, in our case, that it is a real ethical commitment to grant to science the maximum of freedom compatible with the respect due to the other values involved. Hence, the *freedom of science* is part of the ethical consideration of science.

We have thus recognized that the legitimacy of explicitly establishing norms regulating scientific activity cannot be denied. After all, we are already accustomed to the existence of norms regulating pure and applied research from the point of view of security or of secrecy, and one does not see why norms of a more general moral character should be excluded.

However, this fact still leaves open the problem of which agency should be entitled to dictate these norms, and of the way of controlling their application. Following the logic of our discourse, our opinion is that these norms should express the necessity of a systems-theoretic harmonization of different values¹, and hence be the result of a multilateral assumption of *responsibility*: the responsibility of the scientific community towards other values that are present in society, and the responsibility of other social agencies (economic, political, religious, etc.) towards the rights of science. This appeal to responsibility, moreover, is the most appropriate to express the genuine character of any ethical attitude, since responsibility implies at the same time *freedom* and *obligation*, but an obligation which is not equivalent to *compulsion* or *imposition*. To reach this stage of responsibility, a process of maturation, education and *participation* is needed, which implies that scientists should become more sensitive to the existence and significance of more universal human values, by participating in the discussions concerning them and deepening their understanding of their nature and the conditions for their fulfilment. This also means that moralists, theologians and politicians should also become more sensitive and better acquainted with the real issues involved in the practice of scientific research.

It follows from this that a pure and simple self-regulation of the scientific community might not be sufficient, and that some legal regulation, expressing the result of the said mutual collaboration and understanding, would be appropriate. On the other hand a reasonable flexibility should characterize this regulation, except for a few very specific and exactly described cases of particular gravity. The respect of the norms concerning these specific cases should be controlled through the usual means adopted by any public authority for controlling the implementation of laws, while the more flexible norms should be subject to the mechanisms that are usually prescribed by the deontological codes of the different professions.

The impact of science on ethics

What we have said concerning the cooperative spirit which should inspire the establishment of ethical and legal regulations of science does

¹See AGAZZI, E., 1987, *A Systems-Theoretic Approach to the Problem of the Responsibility of Science*, Z. allgemeine Wissenschaftstheorie XVIII (1-2), pp. 30-49.

not simply express the obvious need of some “democratic” way of solving this urgent problem, but corresponds to a much deeper understanding of the relations between science and ethics, an understanding which has again to do with the systems-theoretic view mentioned above. In fact, in speaking of the relationships between science and ethics, it is insufficient to consider the influence that ethics has to exert upon the doing of science, as we have mainly done thus far. An equally interesting investigation concerns the influence of science upon the elaboration of ethics and moral norms. We shall confine ourselves to mentioning here only a couple of examples. Ethics makes use of certain fundamental concepts such as freedom, normality and human nature, and it is clear that a concrete specification of these concepts, and of their *applicability* to actual human actions, requires taking into account the results of several sciences, especially of those concerning man, from biology, to genetics, neurobiology, psychology and sociology. Without correct information being taken from these sciences, it is possible that the ethical discourse be incapable of speaking to the man of today, who has derived from science a new “image” of himself, and thus may be led to feeling that ethics is something obsolete and backwards.

As to the formulation of moral norms, the progress of science (particularly of applied science) has already created and will certainly continue to create quite new and unexpected situations, to which the existent moral norms can hardly apply; or, by suddenly opening unforeseen possibilities of action, and therefore of choice, this progress gives moral relevance to situations which in the past totally escaped the possibility of human decision.

All this indicates that the growth of science imposes a dynamic aspect on morals, which does not mean moral relativism, but making morals capable of coping with the actual situation of contemporary man. This, as we have said, is a consequence of the systems-theoretic approach mentioned above: if morals in general express the imperative of “doing what is right”, without the contribution of other fields they cannot answer the question “*what* it is right to do”, when it comes to concrete situations. Science, without pretending to answer this question (which is not a scientific question), can nevertheless be of help in elaborating the answer.

This Page Intentionally Left Blank

IS THERE ANYTHING WE SHOULD NOT WANT TO KNOW?

PETER GÄRDEFORS

Department of Philosophy, University of Lund, Sweden

1. Introduction

When the wife of the Bishop of Worcester first heard of Darwin's theory of evolution through natural selection, her immediate reaction was: "Descended from monkeys? My dear, let us hope that it is not true! But if it is true, let us hope that it not become widely known!" Her words indicate that she thought that Darwin's theory, if validated, would be dangerous knowledge which should be kept secret among those who could bear the awful truth.

Today, we can smile at her reaction. Darwin's evolutionary theory does not seem dangerous any longer. However, there are a number of present-day examples of scientific investigations that may lead to undesirable knowledge. Let me only remind you of the research on recombinant DNA and the development of ever more powerful nuclear weapons. Until rather recently, it has been taken for granted that all kinds of scientific knowledge are valuable resources, and, consequently, that there should be no restrictions on the topics and directions of scientific inquiry. The question I want to focus on is: *Could there be scientific knowledge, the possession of which would be inimical to ourselves or our welfare?* If the answer is positive, it is natural to argue that we should impose restrictions on scientific research that may produce this kind of knowledge.

Critics of unlimited scientific freedom argue that certain types of knowledge should be forbidden because it is dangerous. But what is meant by "dangerous" here? I want to distinguish between two main types of dangers. The first type is knowledge that will lead to undesirable *material* consequences. The most common fear is that certain forms of

new knowledge will inevitably lead to *technology* that can be misused if put in the wrong hands. Thus knowledge is marked dangerous because it is believed to have dangerous practical consequences.¹ The second type of danger is that certain knowledge is dangerous because of its *mental* consequences. For example it may threaten the established society or some of its institutions or even the established view of humanity itself. Thus some knowledge is thought to be *counter-ideological* or subversive and thus not desirable.²

This distinction between two types of dangers is, admittedly, very rough. Nevertheless, I believe that the distinction is clarifying when attempting to answer the question whether there is anything we should not want to know. The next section will be devoted to a discussion of the problems of inevitable technology in relation to the freedom of scientific investigations. Section 3, then, treats the problems of ideologically unwanted knowledge. As will be clear, rather different considerations will be relevant for these two groups of problems. On the basis of an analysis of some examples of allegedly undesirable knowledge, my conclusion is that *there is no scientific knowledge that we should not want to have*. The final section presents a view on the goals of science, inspired by Aristotle, which supports this conclusion.

Before I start I want to emphasize that I am only concerned with possible restrictions on scientific *knowledge* itself. It is not my aim to discuss restrictions on *methods* used by scientists for obtaining new knowledge. I take it for granted that knowledge is not pursued at all costs, but there will be ethical and other restrictions on what can be done to human and animal subjects, on risky experiments, on the expenses of the research, etc. However, I believe that the problems of potentially undesirable knowledge can and ought to be treated independently of such restrictions on scientific methods.

Another caveat is that I only want to discuss *scientific* knowledge. It is quite a different issue to determine whether there are personal forms of knowledge that we do not want to have. For example, if a doctor discovers that I have a terminal illness, it does not seem obvious that he should tell me about it. I have no exact definition of what constitutes scientific knowledge, but in this context it is sufficient to note that such knowledge is *general* and *impersonal*.

¹ Cf. GRAHAM (1979) on "inevitable technology".

² Cf. BALTIMORE (1979) on the "necessity of freedom".

2. Is any knowledge dangerous for technological reasons?

Fundamental research in various disciplines open up for new and unforeseen technological applications. To take the most dramatic example, fundamental physical research by Einstein and others in the first four decades of this century provided the knowledge necessary for the technological development of nuclear weapons. Some people wish that this knowledge had never been obtained. Shifting to our present situation we may ask whether there are any areas of the frontiers of science that are likely to produce knowledge which leads to undesirable technological consequences; and if so, whether such research should be constrained or prohibited.

I want to present a couple of examples of research that has been claimed to be undesirable in order to determine whether it is the knowledge in itself that is dangerous. SINSHEIMER (1979: p. 29) cites current research upon improved means for isotope fractionation using sophisticated lasers as an investigation of dubious merit. The reason is that "... the most immediate application of isotope fractionation techniques would be the separation of uranium isotopes" and so "... if we devise quick and ingenious means for isotope separation then one of the last defenses against nuclear terror will be breached. Is the advantage worth the price?"

This is a rather typical example of the kind of knowledge that is of "dubious merit" because it leads to undesirable technologies. My main problem with examples of this kind is that it is not obvious that it is the *knowledge* itself that is undesirable. It is clear that if we knew how to separate isotopes efficiently, we could easily find a number of very useful applications. Sinsheimer himself acknowledges this: "To be sure, there are benign experiments that would be facilitated by the availability of less expensive, pure isotopes. For some years I wanted to do an experiment with oxygen-18 but was always deterred by the cost". Rather it is the technology associated with the use of the knowledge for the separation of uranium-235 that frightens Sinsheimer. Thus, in this example, as in so many others, it is not the knowledge *per se* that is dangerous, since it may be used for beneficial purposes as well, but the use of the *technologies* it makes possible.

My second example is also borrowed from SINSHEIMER (1979: pp. 30–31) and concerns research on the aging process. The objective of this type of research is to understand and ultimately control the processes at play in aging and death. Successful research would lead to a substantial

extension of the normal life span (for an overview of the current status of this type of research and its problems, see MORISON (1979)). Sinsheimer presents his misgivings about research on aging as follows: "Obviously, as individuals, we would prefer youth and continued life. Equally obviously, on a finite planet, extended individual life must restrict the production of new individuals and that renewal which provides the vitality of our species. The logic is inexorable. In a finite world the end of death means the end of birth. Who will be the last born?" (p. 30).

I find it difficult to see that Sinsheimer's argument against this particular type of research is valid. The same line of reasoning can be applied, though with less force, to almost any of kind of medical research directed towards preserving human life. Admittedly, there is a problem, the problem of *overpopulation* on earth, but we have to face this problem anyway, quite independently of current research on the aging process. Again, research on aging does not lead to knowledge we do not want to have; the problem is rather how we shall utilize the knowledge.

On the basis of these two examples, let me turn to a more fundamental question: Why do we want to have new knowledge at all? The most immediate answer is: To *understand* and *control* the world around us. New knowledge often makes it possible for us to control a new factor of our environment (or ourselves), which formerly was in the hands of "Nature", or control an old factor to a higher degree. Having control in turn means having greater power to satisfy one's needs and desires.

To explain more precisely in what way new knowledge can improve our understanding of the world let me present a technical result from decision theory due to GOOD (1967). Suppose a decision maker faces a choice between the alternatives a_1, a_2, \dots, a_n in a particular decision situation. His state of knowledge in the decision situation is expressed by whatever possible states of nature s_1, s_2, \dots, s_m he finds relevant to the decision. Given the estimated probabilities $P(s_i)$ of the states, traditional Bayesian decision theory recommends that the alternative that has *maximal expected utility* be chosen (or one of these alternatives, if there are several).

Next, assume that it is possible for the decision maker to obtain new knowledge about the true state of the world, for example by performing an experiment. Suppose that the possible contents of the new information, e.g. the possible outcomes of the experiment, are e_1, e_2, \dots, e_k . The new probability of state s_i given that the outcome e_j is observed will be $P(s_i/e_j)$. The question is now under what conditions the decision maker will be better off after obtaining the information.

Good's theorem says that if the cost of obtaining the new information is negligible, then the expected value of the alternative chosen after

taking the new knowledge into account will always be larger than the value of the alternative chosen without considering the new information. More precisely (but perhaps more confusingly), the expected value of the maximal expected value of the alternatives in the decision situation including the new information is always at least as large as the maximal expected utility of the alternatives in the original decision situation, and it will be larger as soon as the new information is not statistically irrelevant to all the possible states of nature.³

Note that this result does not imply that the maximal expected utility after the new knowledge is obtained will always be at least as large as the maximal expected utility in the situation before the information is obtained, but only that the *expected* maximal expected utility will be at least as large. Thus it may happen that when the new information is gathered, the maximal expected utility of the alternatives in the given decision will become smaller than before.

If I may take the liberty of generalizing Good's technical results to the more general problem of when new knowledge is beneficial, the result says that the expected value of new knowledge is never negative. In other words: in the long run we will make more rewarding decisions if we try to base them on as much knowledge as possible. Of course, this expected gain must be balanced against the cost of acquiring knowledge.

However, it may happen, and sometimes does happen, that the outcome of an investigation provides us with knowledge that makes our prospects look *worse* than before. On the other hand, we can then conclude that our earlier more positive expectations were based on an overly optimistic picture of the state of the world. The new knowledge makes it possible for us to adjust to a more realistic level of aspiration.

Bayesian decision theory, which is presumed for Good's result, is of limited applicability in decision situations where some of the possible consequences are catastrophic and where knowledge about the relevant probabilities is *unreliable*. The reason is that small differences in the estimated probabilities may have drastic consequences for which alternative is recommended by the principle of maximal expected utility. Here decision makers tend to violate the principle and act cautiously so that they are prepared for a risk that is higher than what is predicted by the available knowledge.⁴

If this cautious approach in decision making is transferred to scientific

³ Cf. GÄRDENFORS (1979) for a more detailed discussion of this result.

⁴ Cf. GÄRDENFORS and SAHLIN (1982) for an analysis of decision making with unreliable probabilities.

research strategies, a consequence is that research in high-risk areas should proceed in *small steps*. This is in contrast to normal scientific practice where one tries to design experiments so that a maximal amount of information can be gained. However, in high-risk areas where the estimates of the probabilities of catastrophes are unreliable it seems preferable to perform only experiments where one has a reasonable theoretical basis for predicting the possible outcomes.⁵

There is, however, a much more immediate way in which new knowledge will improve decision making: new knowledge often opens up a new spectrum of decision *alternatives*; we obtain new tools for acting that were not available before. And, of course, having more alternatives to choose from automatically means that the value of the decision will increase (or at least not become smaller). It is also clear that science is our main way of extending our action potentials. As the microbiologist and Nobel Prize winner BALTIMORE puts it (1979: p. 42): "The new ideas and insights of science, much as we may fight against them, provide an important part of the renewal process that maintains the fascination of life. Freedom is the range of opportunities available to an individual—the more he has to choose from, the freer his choice. Science creates freedom by widening our range of understanding and therefore the possibilities from which we can choose."

Thus, decision theory can help us justify why we want to have new knowledge of all kinds. Are there any other justifications?

At the very beginning of his *Metaphysics*, Aristotle writes: "All men by nature desire to know." The Aristotelian view of the goals of man will be discussed in Section 4. The modern understanding of the nature of man is to a large extent influenced by Darwin's *theory of evolution*. It seems clear that since new knowledge helps us form more rewarding decisions, the desire to know is evolutionarily beneficial. The unique capability of humans to acquire and store (remember) new information has made it possible for our species to adapt to a great variety of ecological conditions. Thus it may be expected that our desire to know is to a considerable extent *genetically* determined.

However, our capacities of knowledge acquisition have, during the rapid development of the last centuries, far surpassed what can be controlled by evolutionary mechanisms. Indeed, nuclear weapons give us the power of exterminating the entire human species within a few hours. Furthermore, the recombinant DNA research shows that we can now

⁵ Cf. FØLLESDAL (1979: pp. 407–408) on "reduction of the risk".

deliberately alter the genetic constitution of ourselves (and of other species). This means that we can no longer use the theory of evolution as a justification for an unrestrained hunt for new knowledge, but we must rely on *ethical* values outside of the context of evolutionary explanations.

I have used arguments from both decision theory and the theory of evolution to show that man's desire to know derives from his desire to control. The focal question of my discussion is whether there is anything that we do not want to know, but a related problem worth considering is whether there are factors which we do not want to *control*, even if we know how. Here, I think it is easier to find cases where a positive answer is motivated. For example, it is relatively easy to determine the sex of a foetus using amniocentesis. Coupled with the contemporary acceptance of abortion, this means that we *can* control the sex of our children. However, this power is not used for this purpose on any detectable scale.⁶

Another, perhaps less clear, class of examples concerns our efforts to control the *environment*. As the use of DDT and other substances have taught us, it is extremely difficult to foresee the *side effects* of tampering with environmental variables — the ecological system is very complex and our knowledge is not sophisticated enough. Some even claim that the ecological system is so complex that it is beyond our powers to control it and, consequently, we should not even try to interfere with the course of nature. A parallel argument has been launched against research into recombinant DNA; the claim is that the genetic code is so complex that we cannot predict, let alone control, the side effects of genetic manipulations.

It is important to note that even if there are clear cases of factors we do not want to control, this does not entail that there is anything we do not want to *know*. On the contrary, the more we know (for instance about the complexity of the ecological system), the better can we foresee cases where our attempts to control will fail.

Let me now return to the question of whether there is anything we do not want to know because it leads to dangerous technology. Even if the arguments from decision theory and evolutionary theory presented above are not decisive, they indicate that knowledge is a valuable asset. Furthermore, I have not been able to find any clear examples where knowledge itself is dangerous. It seems always to be the associated technology that carry the threats. I do not believe that there is a sharp dividing line between fundamental and applied science, but the possibili-

⁶ Cf. MORISON (1979: pp. 213–215).

ties of arguing against the desirability of knowledge become smaller the more fundamental the knowledge. My conclusion is that no types of knowledge or search for fundamental scientific knowledge should be forbidden because of technological threats.

There is also another reason why we should not try to regulate fundamental research because of technological dangers. The reason is that major breakthroughs in science cannot be predicted, let alone programmed. BALTIMORE writes (1979: p. 43): "So if you wanted to cut off an area of fundamental research, how would you be able to devise the controls? I contend that it would be impossible".⁷ Again, the degree of unpredictability of a particular area of research is strongly correlated with how fundamental it is. If one accepts that science exists, one is obliged to accept the surprising and discomfoting results along with the more immediately useful knowledge.

The upshot is that controlling fundamental research is not only unwanted but well nigh impossible because it presumes that we can foresee major breakthroughs. In the cases where we suspect that research will lead to dangerous technologies, we should try instead to control the technology. This may be extremely difficult, but it is not impossible.

Prohibiting research is not the only way of controlling the activities of scientists, but in practice the most efficient control is *fund allocation*. I have no general recommendations for the complex problem of how funds ought to be distributed. Let me only note that the unpredictability of fundamental research makes it extremely difficult to assess the pragmatic (economic or social) value of such research. Consequently, deciding which proportion of research funds should be reserved for fundamental research must be determined on non-pragmatic grounds.

Giving scientists who are working with fundamental research complete freedom as to what they may investigate does not entail that they have no *responsibility* for the outcomes of their research. On the contrary, they will be among the first to *foresee* the possible technological applications of

⁷ In relation to the example concerning research on aging, which was discussed above, he writes: "In such an area of science, history tells us that successes are likely to come from unpredictable directions. A scientist working on vitamins or viruses or even plants is just as likely to find a clue to the problem of aging as a scientist working on the problem directly. In fact, someone outside the field is more likely to make a revolutionary discovery than someone inside the field" . . . "You could close the National Institute of Aging Research, but I doubt that any major advance in that field could be prevented. Only the shutdown of all scientific research can guarantee such an outcome" (1979: p. 43). Also cf. THOMAS (1977: p. 327).

their results. Therefore, if we want to be able to control the technologies made possible by fundamental research, it is extremely important that the researchers *inform* and *warn* us as soon as possible about the conceivable new technologies. A little bit of science fiction will probably do no harm in bringing out the message. The important thing is that new technological possibilities are brought to public attention.⁸

3. Is counter-ideological knowledge dangerous?

Some of the best-known cases of resistance against fundamental research were caused by the new knowledge being in conflict with the established ideals of the ruling powers or their views on the world and man's place in it. Apart from the scientific revolutions that were caused by the new theories, there have been *ideological revolutions* which probably have had greater impact on society in general. In this section, I will start by a brief description of a couple of classical cases before I turn to some contemporary examples of science that are threatening the prevailing ideology.

As a first example, let us consider the Copernican system. This system, especially as espoused by Galilei, was in conflict with the dogmas of the Catholic Church (as well as those of Calvinism) and threatened the authority of the Catholic leaders. It taught that the Earth was not the center of the universe, but only a planet among others. Two consequences of this theory seemed especially dangerous to the Church. The first was that it meant that it was impossible to maintain a principal distinction between the phenomena on the earth and those of the heavens – they were all of the same kind. The second, and perhaps more important, was that Copernicus' system was demeaning to the place of man—the Earth had no particular status among the planets in the system.

Darwin's theory of the origin of the species was also thought to be dangerous to the teachings of the Church. A typical example is the reaction of the wife of the Bishop of Worcester presented at the beginning of this article. Again we find a conflict both with the established world-view (all species have been created by God and remain unchanged since then) and with the view of the status of man (man is God's image

⁸ See PETERSSON (1980) for a discussion of the responsibility of scientists working on recombinant DNA.

and not an animal among others). Again, the new knowledge seemed to entail a denigration of the uniqueness of man.

This is not the place to describe the controversy between Darwinism and various religious doctrines. I only want to point out a change in *scientific ethics* which ultimately derives from the Darwinian revolution. Descartes taught that animals are merely automata; sophisticated automata, admittedly, but without a soul and without moral status. According to Darwin, animals are essentially of the same kind as humans. Rather than demeaning the status of man, as the Church feared, this view has led to an upgrading of the animals. Before Darwin, talking about the emotions and feelings of animals was seen as a category mistake — it was like saying that a clock was happy or in pain. But after Darwin, this was a possible area of scientific investigation in which Darwin himself pioneered by his study of the emotional expressions of monkeys (DARWIN 1872). More importantly, acknowledging animal feelings has led to a new view on the use of animals in experiments: we now have rather elaborate rules for how animals shall be treated in scientific investigations and ethical committees supervising the rules; there are even societies for the protection of animals' rights. I see this as a strong sign of an upgrading of our view of animals, not as a sign of a degradation of the status of man as the Church feared.

Among Sinsheimer's examples of research of "dubious merit" we also find one that I think is best classified as counter-ideological. According to him we should avoid searching for contacts with extraterrestrial intelligence. His main concern is that "if such intelligent societies exist and if we can "hear" them, then we are almost certain to be technologically less advanced and thus distinctly inferior in our development to theirs" (SINSHEIMER 1979: p. 29). Elsewhere, he spells out what he believes to be the consequences of this kind of research: "I wonder if the authors of such experiments have ever considered the impact upon the human spirit if it should develop that there are other forms of life, to whom we are, for instance, as the chimpanzee to us. Once it were realized someone already knew the answers to our questions, it seems to me, the impact upon science itself would be especially devastating. We know from our own history the shattering impact more advanced civilizations have upon the less advanced. In my view the human race has to make it on its own, for our own self-respect" (1976: p. 18).

This is an example of knowledge that is not wanted (by Sinsheimer at least) since it is counter-ideological: it threatens our picture of ourselves as the most intelligent beings we know and therefore threatens to

undermine the fundamental ground for viewing ourselves as *responsible* beings.⁹ This anxiety about man's place in nature is, in my opinion, completely misplaced: who is to judge what "inferior" means in this context? Even if an extraterrestrial species turns out to be more "intelligent" than we are, by our present standards, such a species is, by an overwhelming probability, so different from us that an overall comparison is meaningless. Thus, most of what is characteristically *human* will be unique to us and I am convinced that an eventual encounter with an "advanced" species will not lead to a degradation of humanity.

Speaking of "intelligence", this very concept has recently been at the focus of a controversy concerning counter-ideological knowledge. The source of the controversy was that if "intelligence" was measured by the standard tests, certain human races seemed on the average to be considerably less intelligent than others.¹⁰ This led opponents to demand that research on intelligence in general, and inter-racial research in particular, should be prohibited.¹¹

The problem for the intelligence researchers is that it turned out that the standard tests are very sensitive to *cultural* and *educational* differences. The tests have been developed by Western psychologists and, even if not intended, they involve subtle forms of cultural discrimination. For example, Australian aboriginals did very badly on the standard tests involving numerical and verbal abilities. On the other hand, they were far superior to whites in tasks concerning remembering the locations of a large number of items. The result of this and other findings is that the standard intelligence tests are no longer used for cross-cultural comparisons, but only for limited applications. The reason for this is not that those demanding restrictions on intelligence research have won their case. Rather, it has been realized that "intelligence" as measured by the traditional tests is a concept of rather limited utility for psychological research, and of no use at all for sociopsychology.¹² This development would not have occurred if the proposed restrictions on intelligence research had been enforced. What has happened in this case is that

⁹ In Sinsheimer's words: "To really be number two, or number 37, or in truth to be wholly outclassed, an inferior species, inferior on our own turf of intellect and creativity and imagination, would, I think, be very hard for humanity" (1979: p. 30).

¹⁰ Cf. in particular JENSEN (1969, 1972).

¹¹ Cf. BLOCK and DWORKIN (1974-5), in particular pp. 80-99.

¹² BLOCK and DWORKIN'S (1974-5) analysis, which I find devastating for the concept of IQ, is an excellent source of arguments.

instead of a threatening ideological revolution, we have seen a shift in the status of a scientific concept.¹³

As a final example, it should be noted that some of the concerns about “genetic engineering” are caused by the anxiety that this research may provide the possibility of deliberately changing the *personal identity* of some individuals by manipulating their genetic constitution. In a sense, this is counter-ideological knowledge since our picture of ourselves is to a large extent dependent on our notion of a fixed personal identity. At the present state of genetic technology, there are much greater possibilities of altering personality traits by using various forms of drugs than by using genetic engineering.¹⁴ However, these possibilities have not led to any extensive claims for restrictions on drug research. It is difficult to predict what possibilities will open up when we know enough to change polygenic traits that may lead to considerable changes in a person’s emotional reactions or intellectual capacities. I believe, that in the same way as for drug research, the ideological discussion and the associated public opinion will adjust gradually to avoid any serious clashes between ideology and possible research directions.

In summary, the examples that I have presented do not, if I am right, show that we should avoid or restrict counter-ideological knowledge. The examples have been chosen because they have been claimed to produce knowledge that we do not want to have. Refuting some alleged counter-examples, as my aim has been here, does, of course, not prove that there could not be counter-ideological knowledge that we do not want to have. However, the burden of proof lies, I believe, with those who claim that a particular type of research should be forbidden. On this point, Bok (1979: p. 120–121) writes: “Questions may arise about how and when to fund such [counter-ideological] research; individuals may or may not wish to participate in it. But to forbid research on the basis of such nebulous worries is not only unwise, but doubly illegitimate: it interferes with the liberty of investigators without adequate grounds; and it thereby interferes with the public’s right to know.”

¹³ BLOCK and DWORKIN (1974, 1975: pp. 95–99) argue (with some justification) that there should be a temporary ban on research in racial differences of IQ in the U.S. because *mass media* distort and misinterpret the “results” of IQ research and ignore the methodological problems of the area. However, they are careful to point out that they are not requiring that this kind of research be totally prohibited.

¹⁴ Cf. FØLLESDAL (1979: p. 411).

4. An Aristotelian view of the goals of science

When pondering upon man's desire for knowledge and whether the search for new knowledge should go on without restrictions, one may relate it to the more general question of what is the ultimate *goal* of all man's activities. One popular answer to this eternal problem is that man wants to achieve *happiness*. If we take happiness in the utilitarian sense and try to apply this idea to the goals of scientific research, I believe the result would be something quite different from what we actually see. Probably research would be directed, to a large extent, towards developing a harmless drug that would make us all feel constantly happy.

Even if my prediction concerning the aims of a utilitarian science is totally wrong, I do not believe that the overriding goal of scientific research is to maximize happiness in the utilitarian sense. For an alternative view, I would like to return to Aristotle and his writings on the goals of man.

According to the *Nicomachean Ethics*, happiness does not lie in amusements or bodily pleasures but in virtuous *activities*. Aristotle's basic conception of human activity is that a person who does not exercise his faculties to their fullest capacities, be it in art, politics, or science, is not being realized fully as a human being. Happiness is activity in accordance with virtue and the highest virtue is the contemplative activity (1177^a 10–19). Furthermore, he sees contemplation as an activity desired in itself, not for the sake of something else.¹⁵

What is important here is that Aristotle recognizes a natural hierarchy of faculties, where reason is the most superior, and that happiness, what man strives for, must consist of activities that fully use our faculties. "... therefore, the life according to reason is best and pleasantest, since reason more than anything *is man*" (1178^a 6–7). The quest for knowledge is a part of human nature that should not be denied. This desire to know is an end in itself.¹⁶

¹⁵ "... the activity of reason, which is contemplative, seems both to be superior in serious worth and to aim at no end beyond itself (and this augments the activity) ..." (1177^b 18–20). "... This will be the complete happiness of man, if it be allowed a complete term of life ..." (1177^b 23–24).

¹⁶ In *Metaphysics*, Aristotle writes: "... they philosophized in order to escape from ignorance, evidently they were pursuing science in order to know, and not for any utilitarian end. And this is confirmed by the facts; for it was when almost all the necessities of life and the things that make for comfort and recreation had been secured, that such knowledge began to be sought. Evidently then we do not seek it for the sake of any other advantage; but as the man is free, we say, who exists for his own sake and not for another's, so we pursue this as the only free science, for it alone exists for its own sake" (982^b 20–28).

In more recent times we find similar ideas in the Enlightenment. Kant terms the dictum “*Sapere aude!*”, i.e. “Dare to know!”, the essential idea of the Enlightenment.¹⁷ This idea contributes to the martyr image of Galilei.

Most politicians today have a much more pragmatic attitude towards fundamental scientific research; it does not exist (and should not be allowed to exist) only for its own sake, but also for the good of society, which, according to the politicians, is the highest goal. However, even if it is accepted that there are values that are higher than the acquisition of knowledge, this does not entail that there are things we do not want to know. On the contrary, the decision theoretic arguments presented in Section 2 suggest that new knowledge will help us make better decisions, no matter what principles we use to evaluate the possible outcomes of the decisions.

Above I have considered a number of examples of scientific research that have been claimed to be of dubious value. I have divided them into two classes: knowledge leading to dangerous technologies and counter-ideological knowledge. My conclusion as regards knowledge that may lead to dangerous technology was that it is not the scientific knowledge itself that is dangerous, but the dangers depend on the knowledge being used for certain abominable applications. Furthermore, it is well nigh impossible to restrict new scientific knowledge from being found, while we have at least some tools for preventing the most dangerous forms of technology from being developed.

Most of the examples of counter-ideological knowledge are or have been perceived as dangerous because it threatens our picture of ourselves. However, following the Aristotelian view outlined above, if you are afraid of changing your self-image, then you are afraid of developing your full human potential.¹⁸

¹⁷ See KANT (1963: p. 3).

¹⁸ The following quote from Lewis Thomas is in the same vein: “Is there something fundamentally unnatural, or intrinsically wrong, or hazardous for the species, in the ambition that drives us all to reach a comprehensive understanding of nature, including ourselves? I cannot believe it. It would seem to me a more unnatural thing, and more of an offense against nature, for us to come on the same scene endowed as we are with curiosity, filled to the overbrimming as we are with questions, and naturally talented as we are for the asking of clear questions, and then for us to do nothing about it, or worse, to try to suppress the questions. This is a greater danger for our species, to try to pretend that we are another kind of animal, that we do not need to satisfy our curiosity, exploration, and experimentation, and that the human mind can rise above its ignorance by simply asserting there are things it has no need to know” (THOMAS 1977: p. 328).

Apart from threatening our self-image, counter-ideological knowledge has been claimed to be detrimental to the stability of society or its institutions. Again I believe that it is more dangerous to force science into a mold created by the ideology of various institutions. Let me quote once more from BALTIMORE (1979: p. 42): “Scientific orthodoxy is usually dictated by the state when its leaders fear that truths could undermine their power. Their repressive dicta are interpreted by the citizens as an admission of the leaders’ insecurity, and may thus lead to unrest requiring further repression. A social system that leaves science free to explore, and encourages scientific discoveries rather than trying to make science serve it by producing the truths necessary for its stability, transmits to that society strength, not fear, and can endure.”

The examples I have encountered have not convinced me that there is knowledge that we should not want to have. Obviously, my arguments are dependent on a particular view of the goals of man. Let me conclude by reminding you that on the temple in Delphi was written “Know thyself!” In ancient Greece not even the oracle of Delphi could have prophesied the extent to which we now know ourselves, for example that we have the ability to deliberately modify our basic hereditary building blocks. Nevertheless, I believe that this device is still valid for the goals of scientific investigations and that following it is a hallmark of humanity.

Acknowledgement

I wish to thank Dagfinn Føllesdal, Bertil Rolf and Nils-Eric Sahlin for helpful comments.

References

- ARISTOTLE, *Metaphysics* (translated by W.D. Ross).
 ARISTOTLE, *Nicomachean Ethics* (translated by W.D. Ross).
 BALTIMORE, D., 1979, *Limiting science: a biologist’s perspective*, in: G. Holton and R.S. Morison, eds., *Limits of Scientific Inquiry* (Norton, New York), pp. 37–45.
 BLOCK, N. and DWORKIN, G., 1974, 1975, *IQ, heritability and inequality*, *Philosophy and Public Affairs* 3, 331–409; 4, pp. 40–99.
 BOK, S., 1979, *Freedom and risk*, in: G. Holton and R.S. Morison, eds., *Limits of Scientific Inquiry* (Norton, New York), pp. 115–127.
 DARWIN, C., 1871, *Descent of Man* (Murray, London).
 DARWIN, C., 1872, *The Expression of Emotion in Man and Animal* (London).

- FØLLESDAL, D., 1979, *Some ethical aspects of recombinant DNA research*, *Social Science Information* 18, pp. 401–419.
- GÄRDENFORS, P., 1979; *Forecasts, decisions and uncertain probabilities*, *Erkenntnis* 14, pp. 159–181.
- GÄRDENFORS, P. and SAHLIN, N.-E., 1982, *Unreliable probabilities, risk taking, and decision making*, *Synthese* 53, pp. 361–386.
- GOOD, I.J., 1967, *On the principle of total evidence*, *The British Journal for the Philosophy of Science* 17, pp. 319–321.
- GRAHAM, L.R., 1979, *Concerns about science and attempts to regulate inquiry*, in G. Holton and R.S. Morison, eds., *Limits of Scientific Inquiry* (Norton, New York), pp. 1–21.
- JENSEN, A.R., 1969, *How much can we boast IQ and scholastic achievement*, *Harvard Educational Review* 39, 1–123.
- JENSEN, A.R., 1972, *Genetics and Education* (Harper and Row, New York).
- KANT, I., 1963, *What is Enlightenment?*, trans. Lewis White Beck, in Kant, *On History* (Bobbs-Merrill, Indianapolis).
- MORISON, R.S., 1979, *Misgivings about life-extending technologies*, in: G. Holton and R.S. Morison, eds., *Limits of Scientific Inquiry* (Norton, New York), pp. 211–226.
- PETERSSON, I., 1980, *Genetic engineering and the recombinant DNA debate: a case study in the responsibility of the scientist*, *Stockholm Papers in History and Philosophy of Technology*, Report TRITA-HOT-1002.
- SINSHEIMER, R.L., 1976, *Comments*, *Hastings Center Report* 6.
- SINSHEIMER, R.L., 1979, *The presumptions of science*, in: G. Holton and R.S. Morison, eds., *Limits of Scientific Inquiry* (Norton, New York), pp. 23–35.
- THOMAS, L., 1977, *Notes of a biology-watcher: the hazards of science*, *New England Journal of Medicine* 296, pp. 324–328.

THE ETHICS OF SCIENCE AS A FORM OF THE COGNITION OF SCIENCE

BORIS G. YUDIN

Institute of Philosophy, Moscow, USSR

Today we are witnessing an unprecedented upsurge of public interest in the motley and loose arrangement of problems called the ethics of science. This sphere invites, at the same time, an ever growing number of research projects and heated debates.

My aim here is to demonstrate that in the same way as the name of this discipline presupposes two types of cognition, practical research in this field is effected in two-dimensional space.

The dichotomy of what ought to be and what exists is the linchpin of ethical research: our efforts are aimed at substantiating what ought to be and at evaluating what exists from the point of view of what ought to be. Science rests on the dichotomy between theory and practice. Far from embracing the entire range of scientific activities and notions, this dichotomy, nevertheless, is important for our idea of science. I would like to note here that both dichotomies share a certain structural affinity: they concern themselves with the ideal and the really existing. On more than one occasion this situation has caused confusion both in ethics and in science. The ethics of science as a blend of the two spheres, is prone to such confusions to an even greater degree.

In more general terms, the difference between what exists and what ought to be, on the one hand, and theory and practice, on the other, is clear enough: they are two different types of delineating between the real and the ideal. I will not go into greater detail here. I feel it really important to note that research in the ethics of science combine, to varying degrees, two components of different types. It is also advisable to look closer at everything which relates to the ethics of science as a form of science of science.

With this aim in view it is expedient to distinguish (with a full

realisation of this operation's conventional nature) between *discussions* on the ethics of science and *research* in this field. Discussions unfold around specific issues (such as experiments with recombinant DNA or biomedical research). They may touch upon nuclear power or psychotropic preparations. Most general issues, some going back to the age of the Enlightenment, trigger heated debates. One such issue is whether advance in science contributes to the moral perfection of man and humanity as a whole. Such discussions in essence aim at defining the correlation between what ought to be and what exists. Scientific advance today (or in the future) has radical changes in store for mankind. These changes should be assessed from the standpoint of what ought to be; they can also make our ideas of what ought to be more precise and specific.

As in ethics, the ethics of science naturally appraises newly acquired knowledge and practical steps as being desirable or undesirable, morally acceptable or justified. In short, it operates within the range of modalities that belong to the sphere of moral judgments. This determines one of the two dimensions of the two-dimensional space (referred to above) and provides an impetus for the development of the ethics of science. Today, the research impetus comes to the fore in works of another genre. This enables us to regard the sphere of the ethics of science as the two-dimensional space formed by the axes of moral and cognitive judgments.

In what sense can one speak of the ethics of science as a field of research as well as moral evaluations? Contemporary research in ethics proper is mainly analytical and concentrates on the logical structure and typology of moral judgments. Evidently, this approach is too general to bring out the specific content of the ethics of science since, logically speaking, the moral judgments born in this field have nothing to distinguish them from moral judgments formed in other fields.

The ethics of science is not concerned, as a rule, with either formulating a moral system or setting moral norms. It uses the already existing norms and systems. Highly illustrative in this respect was the discussion on the ethical problems of biomedical research and clinical experiments which took place at the 7th International Congress of the Logic, Methodology and Philosophy of Science (Salzburg). HARE (1986) noted that an ethical analysis of clinical experiments brings two moral systems, the utilitarian and the intuitive (deontological) approaches, into collision. The utilitarian approach justifies an experiment if positive expectations are greater than the harm it might cause. This is a controversial approach since it disregards the rights of those people on whom the experiment is being conducted: the informed consent of the patient or (in case of a

child, for instance) his guardian; the right to control the information relating to his own person; the right to be properly treated (patients from control groups are often deprived of this right), etc. This situation, hold the intuitionists, runs contrary to the commonly adopted moral precepts and should, therefore, be rejected.

Having analysed the problem, Hare reaches the conclusion that these approaches should be combined. At the first (initial) level, just as in everyday life, we are to be guided by our intuition. In those spheres, however, where the moral principles clash or where there are no clear-cut moral principles (in short, where our intuition fails us) we have no choice but to switch to the utilitarians' position.

What is even more suggestive (when we have in mind the research trends in the ethics of science) is that Hare bypasses the question of whether each particular moral system is justified. In this respect his work belongs to *applied* ethics.

TRANØY (1986) opposes Hare with a no less interesting proposition. He emphasises that regardless of the importance of the distinction between the deontological arguments (based on the moral norms, rules, rights and duties) and the utilitarian approach (which looks at the results rather than actions and which operates with axiological concepts, with the ideas of the good and evil, etc.), this distinction is of purely academic interest. It attracts philosophers of ethics; those who are confronted with the issues of the biomedical ethics as a matter of routine have no use for abstract discussions on this theme. I find this view to be an adequate description of both the practical approach to biomedical research and to the ethics of science issues.

It is significant that the theme which two and more decades ago attracted defenders and opponents from all camps, has lost its attractiveness. I refer to the idea of changing ethics into rigorously constructed and substantiated scientific knowledge; and to the idea of the ethics of science (based on specific values) as a pattern of ethics in general, etc. (for more details see FROLOV and YUDIN 1986: pp. 111–128). It seems that this situation is a sign of the declining authority of radical scientism and of emergence of the ever new ethical problems within science and its relations with the outside world. We shall discuss this in greater detail below. Here it is pertinent to note that this situation confirmed once more the lack of interest in the problems of substantiation of ethics and the logical analysis of moral categories and judgments on the part of those involved in the ethics of science research.

Let us look at the positive descriptions of the subject-matter of the

ethics of science. The negative description of sorts given above has made it clear that since the ethics of science is not concerned with the problems within the subject-matter of ethics in general, it studies the phenomena occurring in science. One can be even more definite: the ethics of science is a form of *cognition* (or *study*) of science. Or, if we take into account the distinction between discussion and research, it becomes clear that the ethics of science studies precisely science (seen from a specific point of view).

The ethics of science has found its place in the field of the science of science. This is confirmed by the fact that three successive international congresses on the logic, methodology and philosophy of science also discussed the problems of the ethics of science.

Since we recognise the ethics of science as a form of the study of science we admit at least a *possibility* of a corresponding subject of study existing within science. In other words, there exist moral judgments and assessments as applied to science.

This is a purely formal argument; in real life it has many opponents. CHAIN (1970: p. 166) says: "Let me, first of all, state that science, so long as it limits itself to the descriptive study of the laws of Nature, has no moral or ethical quality. . . . The moral and ethical issues, the questions of right or wrong, arise only when scientific research concerns itself with influencing Nature, and this is, of course, next to describing Nature, its major objective. In discussing moral issues. . . we are, therefore, concerned not with descriptive, but applied science. . ."

This is rather a controversial statement. First, influencing Nature as a major objective is seen as subordinate to "the descriptive study of the laws of Nature". In real life, however, another sequence is more frequent or even typical: the objective of describing Nature emerges within the objectives of influencing Nature, of interfering in the natural processes and creating man-made objects and processes. What is described is, as a rule, a situation *created* by the researcher (whether in his mind or in reality).

Chain seems to be rather one-sided in posing the question of the main objective of science as the description of nature or any other objective. Science concerns itself with the cognition of nature which includes not only the description but also the explanation of nature, its understanding, and many other things. Besides, the discussions of the main objective of science, or art, or religion, politics or morals, have no meaning. Who poses these objectives, and to whom? Science as a specific sphere or form of human activity exists and develops irrespective of any predetermined

goal. It is a form in which, and through which, man realises himself, which gives food for culture and for spirit. The people themselves, guided by historical circumstances, determine and follow the goals they pose themselves when they act within the framework of this or other forms.

Chain's position is rather shaky where the question of the goals is concerned. First, the goals always correspond to the means. Would any description of the natural laws justify the destruction of one of the continents? Is there no "moral or ethical quality" in science in this case? The atomic bombardment of Hiroshima and Nagasaki or the Chernobyl disaster gave valuable information about the human reactions and the reactions of other biological species to radiation of various terms. Would this information justify the bombardment and the Chernobyl accident?

No matter whether we are discussing the descriptive or applied science we are inevitably compelled to discuss the means when speaking about goals. This is only natural since goals always exist within the context of *action* and *activity*. Those who reject or restrict (by applied science) the range of moral judgments as related to science overtly or covertly counterpose action and contemplation. This is exactly what Chain does.

One can cite at least two arguments in support of the ethical judgments' relevance *vis-à-vis* science. First, scientific cognition means much more than contemplation, it is activity. AGAZZI (1989) offers ample confirmation of this point. Second, there is every reason to go further than that: the very opposition of activity to contemplation cannot be considered as correct. Essentially, contemplation is a form of activity. As such it is subject to ethical assessments. To be more exact, contemplation as a worthy and even respectable occupation gave rise, in the course of historical development, to science's autonomy and to the possibility to advance through science's own inner criteria. This possibility, writes Agazzi, rests on ethical considerations. Such considerations are necessary to the same degree to which contemplation demands all kinds of resources. The first place belongs to intellectual resources that could be of use in other forms and spheres of activity.

In this way, the ethics of science concerns itself with cognitive activity as an ethically significant activity. What is more, this object can be easily expanded to embrace the entire range of scientific activity because it also includes varied applications of scientific results (in the sphere of ideology and politics, together with technological applications). Scientific activity means sharing knowledge with the scientific community and with the general public (teaching and popularisation of scientific knowledge), and many other things.

There are many different types of activity connected with the production and circulation of science in society which are the privilege of scientists. This privilege makes them responsible for the application and production of scientific knowledge. They are part and parcel of the processes through which new knowledge is obtained. To use a metaphor, they create the field of forces in which scientific cognition unfolds. One may single it out and oppose it to all other types and forms of scientific activity. This operation will, to a considerable degree, be analytical. In real life, however, we are dealing with integral sets of specific scientific activity; each of them can receive ethical appraisal from the point of view of each of the components.

Here is an example. The educational context considerably influences the cognitive activity in science (for more details see PETROV 1981, YUDIN 1986: pp. 186–189). This means that any newly obtained fragment of knowledge should fit into a textbook. The corresponding norms of cognitive activity (which, besides technological, have also ethical content) help realise this condition. This condition itself can be subjected to ethical assessments. It goes without saying that these norms require no deliberation; they are adopted and act, so to speak, automatically. This deliberation becomes necessary when changes occur either in the way the cognitive activity is organised, or in the system of education, or in the way they interact. Such changes may put to question the normative determination and call for its correction or revision.

It looks as if society today faces a crisis in science/society relations: hence the need for the ethics of science. It provides special knowledge about scientific activity and this type of knowledge distinguishes it from all other fields of the science of science. In some way, the current upsurge of ethical research in science reflects the present situation in scientific research. The conceptual schemes and models based on modern scientific materials, however, can be, and are, successfully applied to the history of science. This permits the historians of science to pose new problems in their field and to obtain new interpretations of past events.

I would like to identify three specific features of the present situation in and around science among numerous other features responsible for extended discussion and research into the ethics of science. The first of them is connected with the dramatic increase in the impact of science on the life of man and society. This impact, both real and potential, has intensified: it has become wider and cuts deeper. The increasingly frequent malfunctions of the traditional normative systems which regulate scientific activity are the second specific feature of the present situation.

The third specific feature, changes in society's attitude to science, deserves special attention.

Until recently there existed two clearly delineated attitudes to science: positive and critical, with the positive attitude (an unqualified support for science) predominating. More often than not the normative system and values of science were adopted as standards in other spheres of social life and society as a whole. Anybody who would advise a limitation of these norms and values in cases when they clashed with the norms and values recognised by society would be dismissed as an obscurant.

In the past there were considerably fewer critics of science than today. Even the greatest among them, Rousseau, Dostoyevsky and Tolstoy, were not heeded; their positions were regarded as paradoxes of the great. Both the positive and the critical attitudes were universal, that is, they referred to science as a whole which, undoubtedly, limited their constructive potentialities. A moral sanction applicable to science as a whole was the main object of the controversies concerning science.

In the 1960s came a change in the way public opinion treated science. It was the time when antiscientific views gained wide currency. The nature and the ideological sources of this wide antiscientific movement have received enough attention. However, the fact that the movement evidenced an unprecedentedly keen interest in science and in its social role, has passed unnoticed. This interest lost its academic, abstract nature and became directly involved in moral issues.

Further development resulted in a synthetic attitude which rejects the extremes of both scientism and antiscientism: it can be termed critical-constructive. In more general terms, this attitude does not question the necessity of science and scientific progress. It does not call, however, for a blindfolded approval of science and of the negative results scientific progress entails for man and society. What comes under doubt and what is assessed, from the ethical point of view as well, is not science as a whole but rather specific research trends; what is discussed are possibilities for a timely identification, overcoming and preventing negative results of the advance of science.

The specific features of modern science and its social context make the range of problems of the ethics of science more precise. They also make it possible to define in greater detail the type of knowledge produced by the ethics of science. The distinction between discussion and research formulated above enables us to say that discussions on the ethics of science bring to the fore the critical points in science's advance, the nodes of moral and ethical problems. There are many such points today, each of

them indicating a specific situation which can be resolved only through responsible (from the point of view of ethics) decisions (the variety of such situations is described in detail by FROLOV and YUDIN (1986)). The number of such situations is more than enough to form a *space* of research into the ethics of science.

Far from merely shaping the space of research these critical points also delineate modern science and, especially, its prominent borderline phenomena which are most significant for man. This determines the integral research theme in the ethics of science: what is science? In other words, the ethics of science acts as a theory of scientific activity realised in the extreme and, hence, in some respects, characteristic situations rather than in daily and routine events.

Here a question arises: what defines the place of the ethics of science as contrasted to the methodology of science, sociology of science and the rapidly progressing sociology of scientific cognition?

I think that the question "What is science?" unites all these fields of the science of science. What is more, this question is also relevant for the philosophy of science. For a long time the philosophy of science limited itself by a positive stand *vis-à-vis* science to serving the latter on the technological and, even, axiological planes. There was no doubt about the desirability of scientific progress. The task was to promote it. The methodology of science concerned itself with the ways and means of obtaining reliable scientific knowledge. The sociology of science saw its task in defining social conditions most conducive to scientific progress according to the inner logic of science. As far as the study of the ethical problems of science is concerned, it was dominated by the Mertonian reconstruction of the ethos of science as a system of norms designed to ensure stability of scientific progress. This means that science was maximally free to develop according to its inner logic.

No matter how important these problems are, the methodology, sociology and ethics of science should go farther than that. The wave of discussions (which has been constantly on the rise within the last 15 years) keeps identifying newly recognised critical points in scientific development and the impact of science on man and society. These discussions clearly demonstrate that the question "What is science?" has acquired new dimensions. The philosophy of science of the past ages considered it to be settled, or at least, admitting of a straightforward answer; the task was to formulate the adequate criteria of demarcation between science and non-science (metaphysics, in the first place). This desire to identify

pure science and the emphasis on what distinguishes science from other spheres of mental activity and action are characteristic of the philosophy of science of the preceding periods. It opposed science and other spheres of human thought and action rather than looked for correlations, ties and cooperation.

Today, however, there is a need to elaborate intensively and qualitatively, in many fields, new approaches to the study of science. Thus, the critical points outlined by the discussions on the ethics of science belong not to science alone. Even in cases when discussions probe the deep-lying strata of scientific activity (as was the case in discussions of the limits to experimenting with recombinant DNA), they try to outline the proper sphere of influence of science. This happens, first, because certain aspects of the problems under discussion lie outside science and, second, because these discussions concentrate on the possibility and necessity to limit certain research trends (this problem is debated in the well-known book *Limits of Scientific Inquiry* (HOLTON and MORISON 1979)).

In this way the question "What is science?" occupies the crossroad between science and the human world. The starting point here is the existence of science within this world, not its isolation from it. This leaves the question "What is science?" open. Despite its extremely general wording it presupposes a search for historically specific, rather than abstract-universal, definitions of science.

The openness of this question means that there is no (and cannot be any) predetermined answer to be identified, explained and clothed in suitable wording. The answer is the result of disputes, criticism and self-criticism of different conceptions and views. Each of them is, inevitably, partial and one-sided. It fails to reflect science in its entirety, and modern science especially, as a complex and multifaceted phenomenon. The most important thing is that our searches for an answer presuppose an understanding of the radical changes that science has brought into the world. The very boundaries between science and non-science are not fixed, they change with the course of time. They are, besides, the zones of intensive interaction between science and the phenomena that determine it; this interaction strongly affects both science and the life of man and society.

In these circumstances the answer to the question "What is science?" leaves the sphere of purely academic interest and assumes practical importance. Today the philosophy of science faces the task of comprehending what science is in general and what science is today. This

comprehension outlines the scope of what people expect of science and, consequently, the modern man's position, decisions and actions concerning some of the most significant spheres of his being. What I have in mind here is not prediction of specific scientific results but forecasts of the general trends and general structures of cooperation between science and society and its culture.

Today the concept of science is one of the initial concepts of culture as understood by YUDIN (1978). This is the case because science has assumed an extremely important, many-sided and rapidly expanding role in social life. To enable the culture of today to define itself and to clarify its major problems, we need an analysis of the concept of science (i.e. an answer to the question "What is science?"), realised through the varied manifestations of science. At the same time the efforts of the philosophy of science to answer this question are precisely the efforts to formulate a rational and thoughtful attitude towards science and towards everything connected with it in one way or another.

It is very important that one-sided and narrow answers should not be taken for the final answers and thus block further search. The position occupied by M. Foucault, M. Douglas and their rather numerous French and British followers is prone to this shortcoming. They equate science and power or, to be more exact, regard science as an instrument of domination. Research based on this premise shows many previously ignored aspects of the real existence of science. I think that those who are inclined to absolutise this view, to regard it as the only one rather than an addition to other opinions, commit a grave error. Ours is an age which has demonstrated the negative effects of a one-sided and once popular stand which resembles in many ways the position we are discussing. Here I refer to the interpretation of science as a means of domination over nature.

If we commit ourselves to purely analytical purposes we will probably be able to reduce research activity as a whole, human activity in general and interpersonal relations to the pattern of domination/subordination. This operation will undoubtedly provide a bleak and monotonous picture of reality. Aesthetic considerations apart, this results in stretching the point a great deal. Let us discuss a simple example. I publish an article. This fact can be interpreted as an attempt to impose on my colleagues my own understanding of scientific findings or, even more, my will. This simplified approach fails to explain the multitude of norms (their nature and essence) which guide me in writing the article. Neither it explains the

fact why people resort to similar sophisticated methods to impose their will rather than to use a stick.

The domination/subordination approach cannot explain the distinctions between the desire to impose one's will on others contrary to their own will, the desire to prove one's point with arguments open to critical assessment, or the desire to communicate. Finally, this approach suffers from one more drawback: it passes the struggle science waged against recognised social authorities (there are many examples of this struggle in the history of culture) as an ideological illusion. One can interpret this struggle as the struggle for domination led by social authorities; it should be remembered, however, that this struggle partly promoted the personality type not inclined to blindly accept any authority.

When examining the question "What is science?" as the linchpin of the philosophy of science it is advisable to bear in mind another point of view. Those adhering to it hold that the question is senseless since in reality we are dealing with a multitude of different sciences which have little in common rather than with one science. Suppes defended this view at the Moscow International Congress on the Logic, Methodology and Philosophy of Science. It is not my task to disprove it. I would like merely to point out that in practical terms the search for an answer is more important than the answer itself. The eternal task of the philosophy of science, its debt to culture, is to formulate continually the ever new definitions of science, to criticise and reassess them, to make them more profound and more in line with rapidly changing reality.

I think that this task also determines the place of the ethics of science in the study of science. Though concentrating on certain critical points, the ethics of science is still concerned with science as a whole and not its individual aspects or fields. To be sure, the ethics of science has its own view of science: it provides nothing more than a projection of the three-dimensional phenomenon of science onto the plane. The methodology and sociology of science, like any other field of the science of science, produce their own projections of science. No one of them offers an integral image. To be integral this image requires a combination of all these projections.

This means that the ethics, methodology and sociology of science should cooperate. Significant achievements of this cooperation can be expected not at the level where the ready research findings obtained in one of these fields are applied to other fields. The cooperation which urges constant revision of the initial premises in the study of science and

recognition of their limited and one-sided nature, is much more fruitful. The only road to the multifaceted and integral image of the whole without which we cannot improve our theoretical constructs or (what is more important) to orient ourselves in the world in which science exists, lies in matching all these projections in our imagination.

References

- AGAZZI, E., 1989, *Ethics and science*, in: J.E. Fenstad *et al.*, *Logic, Methodology and Philosophy of Science VIII* (North-Holland, Amsterdam), pp. 49–61, this volume.
- CHAIN, E., 1970, *Social responsibility and the scientist*, *New Scientist*, 48, pp. 166–170.
- FROLOV, I. and YUDIN, B., 1986, *The Ethics of Science: Problems and Discussions* (Politizdat Publishers, Moscow).
- HARE, R.M., 1986, *The ethics of clinical experimentation on human children*, in: R. Barcan Marcus *et al.*, eds., *Logic, Methodology and Philosophy of Science VII* (North-Holland, Amsterdam).
- HOLTON, G. and MORISON, R.S., eds., 1979, *Limits of Scientific Inquiry* (Norton, New York).
- PETROV, M. 1981, *The nature and functions of the processes of differentiation and integration in scientific cognition*, in: *Methodological Problems of Interaction between the Social, Natural and Technical Sciences* (Nauka Publishers, Moscow).
- TRANØY, K.E., 1986, *Experimentation on children: widening the context*, in: R. Barcan Marcus *et al.*, eds., *Logic, Methodology and Philosophy of Science VII* (North-Holland, Amsterdam).
- YUDIN, B., 1986, *The Methodological Analysis as a Trend in Studying Science* (Nauka Publishers, Moscow).
- YUDIN, E., 1978, *The Systems Approach and the Principle of Activity* (Nauka Publishers, Moscow).

1
**Foundations of
Mathematical Reasoning**

This Page Intentionally Left Blank

NON-MONOTONIC REASONING BY AXIOMATIC EXTENSIONS

GERHARD JÄGER

Institut für Informatik, ETH-Zentrum, Zürich, Switzerland

General background

Non-monotonic reasoning is the modern name for a variety of scientific activities that are characterized by the idea that the traditional deductive approach to inference systems is too narrow for many applications and that new formalisms are required which make arrangements for default reasoning, common sense reasoning, autoepistemic reasoning and the like. The recent interest in this field is caused by questions in artificial intelligence (AI) and computer science and has originally led to a series of ad hoc methods and isolated case studies. Gradually, however, something like a theory of non-monotonic reasoning has emerged, and the mathematical and logical foundations of non-monotonic reasoning have been studied.

In the following we will be interested in connections between sets of formulae and individual formulae. We start from a given formal language L and denote the set of all L formulae by For_L . A binary provability relation $\Vdash \subset Pow(For_L) \times For_L$, is called *monotonic* if

$$S \subset T \ \& \ S \Vdash A \Rightarrow T \Vdash A$$

for all sets of L formulae S , T and all formulae A . Otherwise we say that \Vdash is *non-monotonic*. The usual provability relation of classical or intuitionistic logic are monotonic. In the non-monotonic case the number of theorems may diminish by the addition of further axioms.

At first sight non-monotonic provability relations look a little bit strange. However, they are not an invention of this century but have a long history going back to antiquity. Well-known examples are systems of

the so-called *inductive logic* which are used in philosophy of science in connection with the discussion of rise, change and confirmation of scientific theories (cf. e.g. ESSLER 1970).

On a very intuitive level, inductive logic was already studied by ARISTOTELES (1921) and the school of the Epicureans (PHILODEMUS, transl. 1941). These ideas were taken up again and refined in the Renaissance and in the Age of Enlightenment by BACON (1783, 1830) and HUME (1910, 1938). Completely mature systems were introduced at the beginning of this century in the form of CARNAP's *Induktive Logik* (1945, 1958) and POPPER's *Theorie der Bewährung* (1935, 1971).

Inductive logic is historically an interesting example but it does not account for the great success of non-monotonic reasoning during the last few years. Fresh blood was brought into this field by the hope that non-monotonic concepts could be a useful tool for solving central problems of artificial intelligence and logic programming like the design of powerful expert systems and logical data bases. This is a very active research area, and it would not make sense to give specific references. For further details and background information the interested reader should consult the proceedings of the recent conferences on this subject and the relevant journals (e.g. *Journal of Logic Programming, Artificial Intelligence*).

Meanwhile, many different formal representations of the main notions have been introduced and studied, both from the theoretical and practical points of view. A whole line of research has been initiated by the work of McDERMOTT (1982), McDERMOTT and DOYLE (1980), MOORE (1985), REITER (1980a), PARIKH *et al.* (1984) and is basically concerned with the integration on non-monotonic methods in a modal logic framework. Alternative directions avoid modal logic and try to reflect non-monotonicity more directly in first or higher-order logic.

This paper is concerned with forms of non-monotonic reasoning which are induced by what we call *default operators*. Every partial operator

$$H : Pow(For_L) \rightarrow Pow(For_L)$$

induces a provability relation \Vdash_H defined by

$$T \Vdash_H A \Leftrightarrow H(T) \vdash A$$

for all formulae A and all $T \in domain(H)$. \Vdash_H is monotonic if H is a partial monotonic operator from $Pow(For_L)$ to $Pow(For_L)$ with respect

to \subset , but it is also clear that there are many examples of (non-monotonic) operators H such that \Vdash_H is a non-monotonic provability relation; examples will follow later. $H(T)$ is called an *axiomatic extension* of T if it results from T by adding new formulae as further axioms.

The idea of this approach is the following: T is the explicit description of a situation as it is given to us. The transition from T to $H(T)$ reflects the incorporation of default reasoning, and then logical derivability is used.

We will make a further restriction and confine ourselves to default operators that arise in connection with the treatment of negation and negative information in a logic programming environment. In recent years the general theme of *negative information* has attracted a certain amount of attention, especially in the context of logic programming, logical data bases, information processing and the like. LLOYD's textbook (LLOYD 1984) together with articles by CLARK (1978), REITER (1978), SHEPHERDSON (1984, 1985, 1988) and LLOYD and TOPOR (1984, 1985, 1986) provide a very good introduction to the general questions and supply numerous references for further reading.

In the following we will deal with a countable first-order language L with equality and an arbitrary number of function and relation symbols. The terms $a, b, c, a_1, b_1, c_1, \dots$ and formulae $A, B, C, A_1, B_1, C_1, \dots$ of L are defined as usual. Formulae and terms without free variables are called *ground*; ground formulae are often also denoted as *sentences*. We write \underline{a} for a finite string a_1, \dots, a_n of L terms and use the notation $A[\underline{x}]$ to indicate that all free variables of A come from the list \underline{x} ; $A(\underline{x})$ may contain other free variables besides \underline{x} . An L theory is a (possibly infinite) collection of ground L formulae. By $T \vdash A$ we express that the formula A can be deduced from the theory T by the usual axioms and rules of predicate logic with equality. An L theory T is *inconsistent* if every L formula is deducible from T ; otherwise T is *consistent*.

The collection of *Horn clauses* consists of all L formulae of the form

- (i) A
- (ii) $\neg B_1 \vee \dots \vee \neg B_n \vee A$
- (iii) $\neg B_1 \vee \dots \vee \neg B_n$

where A and B_1, \dots, B_n are atomic formulae; Horn clauses of the form (i) and (ii) are called *definite* Horn clauses. If A is a formula, then the universal closure of A is the formula obtained by adding a universal quantifier for every variable having a free occurrence in A . A *logic*

program is a finite collection of universal closures of definite Horn clauses.

In addition, L is supposed to contain a sequence $\underline{P} = P_1, \dots, P_n$ of relation symbols which we assume to be unary in order to keep the notation as simple as possible; the extension of our results to relation symbols \underline{P} of arbitrary arities is straightforward. L_0 is the sublanguage of L without occurrences of the relation symbols \underline{P} .

Before going into further details, let us describe the general scenario. We assume that we are confronted with a huge amount of information concerning a specific field in the scope of our experience (e.g. a particular subject in mathematics, physics, economics, medicine, ...) which may consist of elementary facts, logical dependencies and so on. In the ideal world of logic programming, all knowledge is represented in a formal language; the basic knowledge provides the axioms, and new knowledge is acquired by using logical derivability. However, the amount of information that needs to be represented about some specific domain may be prohibitively large and logically intricate such that questions of computational complexity become very important. One of the most promising approaches to achieve efficiency is the restriction to positive information. Negative information then has to be deduced by using (suitable) forms of default reasoning which often go along with a kind of meta-concepts. This strategy is followed for example in various versions of PROLOG.

More specifically, in the following we will work with the first-order language L and assume that \underline{P} denotes the sequence of predicates to which default reasoning might be applied; our knowledge with respect to the other relation and function symbols is supposed to be stable.

Accordingly, we write every L theory T in the form

$$T = SF + DB_{\underline{P}}(T)$$

where SF is the L_0 theory $T \cap L_0$ and $DB_{\underline{P}}(T)$ the complement of SF in T . Hence T is split into the stable facts SF and the data base $DB_{\underline{P}}(T)$ for \underline{P} .

SHEPHERDSON (1987) presents an excellent survey on negation in logic programming. In this paper we will concentrate on the mathematical and logical aspects of two typical representatives of non-monotonic reasoning that deal with negative information. The first is REITER's closed-world assumption (1978) and the second is MCCARTHY's notion of circumscription (1980). Every approach is based on a different idea but put together they form something like a frame for the present-day activities in this field.

1. Closed-world assumption

Reiter's closed-world assumption *CWA* is the most rigid form to introduce negative information (REITER 1978). The *CWA* is motivated by applications in data base theory and is based on the idea that T contains all positive information and that any positive ground literal which is not implied by T is false. In its original form, the closed world of a theory T is defined as the set of formulae

$$CWA(T) := T + \{\neg A : A \text{ ground atom} \ \& \ T \not\vdash A\}.$$

The relativized version of the closed-world assumption with respect to the relation symbols $\underline{P} = P_1, \dots, P_n$ is given by

$$CWA_{\underline{P}}(T) := T + \bigcup_{i=1}^n \{\neg P_i(a) : a \text{ ground term} \ \& \ T \not\vdash P_i(a)\}.$$

REMARK 1.1. It is easy to see that the default operator $CWA_{\underline{P}}$,

$$CWA_{\underline{P}} : Pow(For_L) \rightarrow Pow(For_L); \quad T \rightarrow CWA_{\underline{P}}(T)$$

is defined for all sets of sentences and induces a non-monotonic provability relation $\Vdash_{CWA_{\underline{P}}}$. To give an example, let SF_3 be a theory which formalizes that its universe consists of the three different elements a , b and c . Then define

$$T_0 := SF_3 + \{P(a)\}, \quad T_1 := SF_3 + \{P(a), P(b)\}.$$

It follows

- (i) $T_0 \subset T_1$;
- (ii) $T_0 \not\vdash P(b)$, i.e. $CWA_{\underline{P}}(T_0) \vdash \neg P(b)$;
- (iii) $T_1 \vdash P(b)$, i.e. $CWA_{\underline{P}}(T_1) \not\vdash \neg P(b)$.

The closed-world assumption is a very handy and well-motivated concept as long as elementary assertions about \underline{P} are considered. Then the meaning of $CWA_{\underline{P}}(T)$ is perfectly clear and its use causes no problems. However, as soon as more complex situations are taken into account, one has to be very careful.

REMARK 1.2. There exist consistent theories T such that $CWA(T)$ and $CWA_{\underline{P}}(T)$ are inconsistent.

Examples of this kind are well known in the literature and a specific example will be presented later. The following theorems treat the problem of consistency of the non-relativized closed-world assumption. The first is due to REITER (1978); the second is proved by MAHR and MAKOWSKY (1983) and MAKOWSKY (1985) and is based on a theorem of MALTSEV (1971).

THEOREM 1.3 (Reiter). *If T is a set of definite Horn clauses without equality, then $CWA(T)$ is consistent.*

THEOREM 1.4 (Mahr, Makowsky, Maltsev). *If for each set S of ground atoms, possibly involving new constants, $CWA(T + S)$ is consistent, then T is equivalent to a set of definite Horn clauses.*

SHEPHERDSON (1988) contains an elementary proof of Theorem 1.4 and an example which shows that this theorem is no longer true if the sets S are restricted to sets of ground atoms in the language of T .

The relativized closed-world assumption is significantly more general and, accordingly, the question of consistency is a different matter. Besides this, we are not so much interested in criteria which refer to theories in extended languages but favour the idea that all considerations should be developed within the syntactic framework (formal language) L which is given in advance.

A natural characterization of those theories T which give rise to a consistent $CWA_{\underline{P}}(T)$ has not yet been found, and it is not clear whether a satisfactory one exists. However, there are partial results which cover most of the relevant cases. Before stating them, we need some further terminology.

DEFINITION 1.5.

(1) An L formula A is called \underline{P} *positive* if A does not contain negative occurrences of the relation symbols P_1, \dots, P_n .

(2) A formula A is called a Σ *formula* [Π *formula*] if all of the quantifiers in its prenex normal form are existential [universal].

(3) An L sentence is called *inductive with respect to \underline{P}* if it is of the form

$$(\forall x)(B[\underline{P}, x] \rightarrow P_i(x))$$

where $B[\underline{P}, x]$ is a \underline{P} positive formula. It is called Σ inductive if, in addition, $B[\underline{P}, x]$ is a Σ formula.

(4) A set of L formulae is called inductive [Σ inductive] with respect to \underline{P} if each of its elements is inductive [Σ inductive] with respect to \underline{P} .

Inductive and Σ inductive sentences are typical candidates for passing on positive information about \underline{P} . They state what is true provided that something else is true, but they do not state what is false. Typical examples of Σ inductive sentences are the usual definition of the natural numbers

$$(\forall x)[(x = 0 \vee (\exists y)(P(y) \ \& \ x = y + 1)) \rightarrow P(x)]$$

and all definite Horn clauses (they are Σ inductive with respect to their relation symbols). However, the following example shows that inductivity of the data about \underline{P} does not guarantee the consistency of the closed-world assumption.

1.6. Example

Let SF be an incomplete L_o theory with two provably different constants a_1, a_2 , and suppose that A is a ground L_o formula such that SF proves neither A nor $\neg A$. We define

$$B(P) := (\forall x)[(A \ \& \ x = a_1) \vee (\neg A \ \& \ x = a_2) \rightarrow P(x)] ,$$

$$T := SF + \{B(P)\} .$$

$B(P)$ is inductive with respect to P . It is also clear that T proves $P(a_1) \vee P(a_2)$ but does not prove $P(a_i)$ for $i = 1, 2$. Using the closed-world assumption we conclude that $P(a_1) \vee P(a_2)$ and $\neg P(a_1) \ \& \ \neg P(a_2)$ are theorems of $CWA_p(T)$. Hence $CWA_p(T)$ is inconsistent.

Our next goal is to find criteria for the stable facts which, together with the inductivity of Σ inductivity of the data about \underline{P} , ensure the consistency of $CWA_p(T)$. To achieve this, one has to make sure that the models of SF are not too different. The situation is extreme for inductive data bases; in the case of Σ inductive data bases we can be more liberal (cf. Definition 1.10).

DEFINITION 1.7. An L_o theory Th is called *weakly categorical* (in L_o) if Th has a countable model, and any two countable models of Th are

isomorphic. (Here a model is denoted as countable if the cardinality of its universe is less than or equal to ω .)

This definition means that the class of all countable models of a weakly categorical L_o theory has, up to isomorphism, exactly one element. It does not say whether there are uncountable models and how many. The concepts “weakly categorical” and the familiar “ ω categorical” (see e.g. CHANG and KEISLER 1973) are related but not identical.

If we ignore uncountable structures, then one can think of a weakly categorical theory as a theory that provides enough information to pin down the universe and the meaning of all function and relation constants. Weak categoricity is a very strong assumption if we deal with theories which have infinite domains. For applications, however, weak categoricity is more important in connection with finite domains. Then there are many examples of weakly categorical theories.

THEOREM 1.8. *Let T be the theory $SF + DB(\underline{P})$ and assume that*

- (A1) *SF is a weakly categorical L_o theory,*
- (A2) *$DB(\underline{P})$ is inductive with respect to \underline{P} .*

Then $CWA_{\underline{P}}(T)$ is a conservative extension of T for all sentences which are \underline{P} positive.

COROLLARY 1.9. *Let T be the theory $SF + DB(\underline{P})$. If SF is a weakly categorical L_o theory and $DB(\underline{P})$ a data base which is inductive with respect to \underline{P} , then $CWA_{\underline{P}}(T)$ is consistent.*

Theorem 1.8 and its Corollary 1.9 are proved by JÄGER (1987). This paper also briefly addresses the question about the converse of Theorem 1.8 and shows that this result is sharp in some sense, at least in the presence of the domain closure property. Domain closure has the effect of guaranteeing that every element of the universe has a name (cf. REITER 1980b).

DEFINITION 1.10. A model \mathbf{M} of the L_o theory Th is called a *primary model* of Th if every model of Th has a substructure which is isomorphic to \mathbf{M} .

Peano arithmetic is a typical example of a theory that has a primary model, namely the standard model of the natural numbers. The notion of primary model introduced here resembles the notions of prime model (see e.g. CHANG and KEISLER 1973) and initial model (see e.g. GOGUEN *et al.* 1977) but is not equivalent to one of these. In the case of prime models, the term substructure is replaced by elementary substructure, and in the case of initiality, uniqueness of the substructure is required.

THEOREM 1.11. *Let T be the theory $SF + DB(\underline{P})$ and assume that*

(B1) *SF is an L_o theory which has a primary model*

(B2) *$DB(\underline{P})$ is Σ inductive with respect to \underline{P} .*

Then $CWA_{\underline{P}}(T)$ is a conservative extension of T for all Σ sentences which are \underline{P} positive.

COROLLARY 1.12. *Let T be the theory $SF + DB(\underline{P})$. If the L_o theory SF has a primary model and the data base $DB(\underline{P})$ is Σ inductive with respect to \underline{P} , then $CWA_{\underline{P}}(T)$ is consistent.*

These results settle the question of the consistency of the relativized closed-world assumption for a large class of theories. In connection with the theorems of Reiter and Mahr–Makowsky–Maltsev for the non-relativized case they provide the ground for a justified use of the closed-world assumption.

Advantages of the closed-world assumption are its clear methodological conception and its efficiency for elementary data base. Disadvantages are the restricted range of applicability and its complicated proof procedure.

Working with the closed-world assumption means working in two different levels. In the first level one has the theory T and checks whether certain atomic sentences $P_i(a)$ are provable or not. In the second level, T is extended by negations of some non-provable atoms, and then the usual derivation procedure is initiated. Formally this is reflected by the fact that provability with the closed-world assumption is Σ_2^0 and not Σ_1^0 as usual.

REMARK 1.13. It is possible to generalize Theorem 1.8, Theorem 1.11 and their corollaries such that the data bases may contain sentences $\neg B(\underline{P})$ where $B(\underline{P})$ is a \underline{P} positive $[\Sigma]$ formula. For details see JÄGER (1987).

2. Circumscription

Induction principles are the mathematical form of introducing negative information. If we state that (i) $2N$ is a set which contains 0 and is closed under addition of 2, and (ii) every set with these closure properties contains $2N$ as a subset, then we implicitly say that, for example, the odd natural numbers do not belong to $2N$. In this sense the inductive definition of a particular set allows one to prove that certain elements do not belong to this set.

McCarthy's notion of circumscription is based on the idea of using adequate modifications of induction principles for the purpose of formalizing default reasoning and common sense reasoning in AI contexts (McCARTHY 1980, 1986). He has actually introduced various forms of circumscription but these versions are (from a logical point of view) more or less equivalent, and in this paper we concentrate on predicate circumscription.

From now on we assume that the data base $DB(\underline{P})$ for \underline{P} consists of a finite set $\{D_1(\underline{P}), \dots, D_m(\underline{P})\}$ of L sentences. Then we can form the conjunction $DB_{\&}(\underline{P})$ of the elements of $DB(\underline{P})$, i.e.

$$DB_{\&}(\underline{P}) := D_1(\underline{P}) \& \dots \& D_m(\underline{P}).$$

It is clear that $DB_{\&}(\underline{P})$ is (equivalent to) a $[\Sigma]$ inductive sentence if $DP(\underline{P})$ is $[\Sigma]$ inductive.

For sequences of L formulae $\underline{F} = F_1, \dots, F_n$ and $\underline{G} = G_1, \dots, G_n$ we introduce the following shorthand notation

$$\begin{aligned} \underline{F} \subset \underline{G} &:= (\forall x)(F_1(x) \rightarrow G_1(x)) \& \dots \& (\forall x)(F_n(x) \rightarrow G_n(x)), \\ \underline{F} = \underline{G} &:= \underline{F} \subset \underline{G} \& \underline{G} \subset \underline{F}. \end{aligned}$$

Let T be the theory $SF + DB(\underline{P})$ where SF is an L_o theory and $DB(\underline{P})$ a finite set of L sentences. The *circumscription* of T with respect to \underline{P} is the theory $CIR_{\underline{P}}(T)$ which has as axioms

- (Cir. 1) all axioms of SF ;
- (Cir. 2) $DB_{\&}(\underline{P})$;
- (Cir. 3) $DB_{\&}(\underline{F}) \& \underline{F} \subset \underline{P} \rightarrow \underline{P} \subset \underline{F}$ for all L formulae $\underline{F} = F_1, \dots, F_n$.

The axiom scheme (Cir. 3) is an induction principle which formalizes that there are no definable proper subsets of the relations $\underline{P} = P_1, \dots, P_n$

which satisfy $DB_{\mathcal{g}}$. Put in other words, it means that P_1, \dots, P_n is a sequence of *minimal witnesses* for the definition clause $DB_{\mathcal{g}}$.

2.1. Examples

Let SF be a theory which contains the constants and defining equations for all primitive recursive functions and relations.

(1) $T_0 := SF + \{P(0) \ \& \ (\forall x)[P(x) \rightarrow P(x')]\}$. Then $CIR_P(T_0)$ proves for all formulae F : $F(0) \ \& \ (\forall x)[F(x) \rightarrow F(x')] \rightarrow (\forall x)[P(x) \rightarrow F(x)]$.

(2) $T_1 := SF + \{P(0) \vee (P(1) \ \& \ P(2))\}$. Then $CIR_P(T_1)$ proves $P = \{0\} \vee P = \{1, 2\}$.

(3) $T_2 := SF + \{(\forall x)[(\forall y \geq x) \neg P(y) \rightarrow P(x)]\}$. Then a subset I of the natural numbers satisfies this condition for P if and only if it is infinite. Since there is no minimal infinite subset of the natural numbers, we can easily conclude, that $CIR_P(T_2)$ is inconsistent.

REMARK 2.2. The example given in Remark 1.1 can also be used to show that the (partial) default operator CIR_P ,

$$CIR_P: Pow(For_L) \rightarrow Pow(For_L); \quad T \rightarrow CIR_P(T)$$

induces a non-monotonic provability relation $\Vdash_{\underline{P}}$. The operator $CIR_{\underline{P}}$ is defined for all theories $T = SF + DB(\underline{P})$ with a finite data base $DB(\underline{P})$ for \underline{P} .

From the experience in proof theory one knows that it is nearly impossible to characterize those theories T which remain consistent under extension by circumscription. Hence the real problem consists of singling out natural classes which preserve consistency and are general enough to include many interesting applications of circumscription. The notion of *positive disjunctive circumscription* is a step in this direction.

DEFINITION 2.3.

(1) For every sequence $\underline{a} = a_1, \dots, a_m$ of and every unary relation symbol R we introduce the following abbreviation

$$\underline{a} \in_d R : \Leftrightarrow R(a_1) \vee \dots \vee R(a_m).$$

(2) An L sentence is called *positive disjunctive with respect to \underline{P}* if it is of the form

$$(\forall \underline{x})(B[\underline{P}, \underline{x}] \rightarrow \underline{x} \in {}_a P_i)$$

where $B[\underline{P}, \underline{x}]$ is a \underline{P} -positive formula.

2.4. Examples

Every inductive formula is positive disjunctive (with respect to the same relation symbols). If $B[P, x, y]$ is P positive, then the formula

$$(\forall x)(\forall y)(B[P, x, y] \rightarrow P(x) \vee P(y)) \quad (*)$$

is positive disjunctive in P .

The concept of positive disjunctive definition extends the notion of positive inductive in a non-trivial way and gives rise to a series of interesting questions. First we state a central lemma which guarantees the existence of minimal witnesses for positive disjunctive sentences.

LEMMA 2.5. *Let $T = SF + \{D(\underline{P})\}$ where $D(\underline{P})$ is positive disjunctive with respect to \underline{P} and $SF \subset L_0$. Assume also that \mathbf{M} is a model of SF with universe $|\mathbf{M}|$.*

(1) *There exists a sequence \underline{I} of minimal subsets of $|\mathbf{M}|$ such that $\mathbf{M} \models D(\underline{I})$.*

(2) *If \underline{J} is a sequence of subsets of $|\mathbf{M}|$ such that $\mathbf{M} \models D(\underline{J})$, then there exist minimal subsets \underline{I} of $|\mathbf{M}|$ with the properties $\mathbf{M} \models D(\underline{I})$ and $\underline{I} \subset \underline{J}$.*

A proof of Lemma 2.5 is given by JÄGER (1986). It consists of a non-constructive argument which is based on a combination of pigeon hole principle and Zorn's lemma (or well ordering theorem as in JÄGER (1986)). Here we make use of this lemma in order to infer the following theorem and its corollary.

THEOREM 2.6. *Let T be the theory $SF + \{D(\underline{P})\}$ and assume that*

- (C1) *SF is a consistent L_0 theory,*
- (C2) *$D(\underline{P})$ is a positive disjunctive L sentence.*

Then $CIR_{\underline{P}}(T)$ is a conservative extension of T for all sentences which are \underline{P} positive.

COROLLARY 2.7. *Let T be the theory $SF + \{D(\underline{P})\}$. If SF is a consistent L_o theory and $D(\underline{P})$ a positive disjunctive L sentence, then $CIR_{\underline{P}}(T)$ is consistent.*

Our next considerations are concerned with the norms of minimal positive disjunctively definable sets. The notion of norm is a central concept in the theory of inductive definitions and (implicitly) used in many fundamental results like prewell-ordering theorem, stage comparison theorem, closure theorem or (recursion-theoretic) boundedness theorem. The theory of inductive norms is systematically developed by MOSCHOVAKIS (1974) and BARWISE (1975) where the emphasis is put on the recursion theory and definability theory of inductively definable sets. In proof theory, inductive norms play an important role in connection with the so-called provable parts of inductively definable sets and proof-theoretic ordinals (see e.g. BUCHHOLZ *et al.* 1981).

Now we extend the notion of norm to sets which are defined by positive disjunctive definitions and follow the classical approach as closely as possible. To make the notation more comprehensible, we restrict ourselves to positive disjunctive sentences of the form (*). The extension of our results to the general case is then a matter of routine.

Let SF be a consistent L_o theory, \mathbf{M} a model of SF and I a minimal witness for the positive disjunctive sentence $D(P)$,

$$D(P) : \Leftrightarrow (\forall x)(\forall y)(B[P, x, y] \rightarrow P(x) \vee P(y))$$

where the formula $B[P, x, y]$ is P positive. Then the set I can be split into stages by using the following recursion on the ordinals:

$$I_{<\alpha} := \bigcup_{\xi < \alpha} I_{\xi};$$

$$I_{\alpha} := \{m \in I : \mathbf{M} \models (\exists x)(\exists y)(B[I_{<\xi}, x, y] \ \& \ (m = x \vee m = y))\}.$$

Since the formula $B[P, x, y]$ is P positive, we have $I_{\alpha} \subset I_{\beta}$ for all $\alpha < \beta$. Hence there exists an ordinal γ such that $I = I_{\gamma} = I_{<\gamma}$. The least such ordinal is called the norm of I and denoted by $\|I\|$;

$$\|I\| := \min \{ \xi : I = I_{\xi} \}.$$

In the definition of the stages I_α we explicitly refer to the set I . The set I must be given first and is then split into stages. This is a significant difference to the case of positive inductive definitions where the stages are defined from scratch and then used to determine the least fixed point of the corresponding inductive definition (cf. MOSCHOVAKIS 1974).

REMARK 2.8. The norm of a minimal positive disjunctively definable set is not an invariant of the corresponding definition. JÄGER (1986) gives a trivial example of a positive disjunctive sentence $D(P)$ which has two minimal witnesses I and I' such that $\|I\| \neq \|I'\|$.

A lot is known about the norms of positive inductive definitions. One of the main results states that, over the standard structure of the natural numbers, the norm of each set which is definable by a positive inductive definition is less than or equal to the first non-recursive ordinal ω_1^{CK} . It seems to be an interesting and open question whether a similar result is also true for positive disjunctive definitions.

2.9. *Open questions* (for $SF =$ Peano arithmetic and $\mathbf{M} =$ standard structure of the natural numbers)

(1) If I is a minimal witness for the positive disjunctive sentence $D(P)$, then we have $\|I\| \leq \omega_1^{CK}$. (??)

(2) For every positive disjunctive sentence $D(P)$ exists a minimal witness I such that $\|I\| \leq \omega_1^{CK}$. (??)

JÄGER (1986) is also concerned with some proof-theoretic aspects of positive disjunctive circumscription. Among other things, we prove a boundedness theorem which establishes the connections between the stages of minimal witnesses for positive disjunctive sentences and provability in $CIR_p(T)$.

As usual, take ε_o to denote the least ordinal ξ such that $\omega^\xi = \xi$. For every limit ordinal λ less than ε_o we then define its fundamental sequence $(\lambda[n] : n < \omega)$ by the following recursion:

- (i) $\omega[n] := n$;
- (ii) if λ is the ordinal $\omega^\alpha + \beta$, where $\beta < \omega^{\alpha+1}$, then $\lambda[n] := \omega^\alpha + \beta[n]$;
- (iii) $\omega^{\alpha+1}[n] = \omega^\alpha + \dots + \omega^\alpha$ (n summands);
- (iv) if λ is of the form ω^α and α is a limit ordinal, then $\lambda[n] := \omega^{\alpha[n]}$.

This covers all possible cases. As a consequence we obtain $\lambda = \sup \{ \lambda[n] : n < \omega \}$ for all limit ordinals $\lambda < \varepsilon_0$. Using these fundamental sequences, we now introduce a hierarchy $(f_\alpha : \alpha < \varepsilon_0)$ of number-theoretic functions from ω to ω :

$$\begin{aligned} f_0(n) &:= n + 1; \\ f_{\alpha+1}(n) &:= f_\alpha(f_\alpha(\dots f_\alpha(n) \dots)) \quad (n \text{ applications of } f_\alpha); \\ f_\lambda(n) &:= f_{\lambda[n]}(n), \quad \text{if } \lambda \text{ is a limit ordinal.} \end{aligned}$$

The hierarchy $(f_\alpha : \alpha < \varepsilon_0)$ is known as the *fast growing hierarchy* and is important for the classification of number-theoretic functions; f_ω corresponds to the Ackermann function. If we work in an extension Th of Peano arithmetic PA , we use (IND_N) to denote the scheme of complete induction

$$(IND_N) \quad F(0) \ \& \ (\forall x)(F(x) \rightarrow F(x')) \rightarrow (\forall x)F(x)$$

for all formulae of the language of Th . For the definition of the *Howard ordinal* η_0 we refer for example to BUCHHOLZ *et al.* (1981).

THEOREM 2.10. *Let $A[\underline{x}, \underline{z}]$ be an arbitrary L_o formula, $D(P)$ a positive disjunctive L sentence, $T = SF + \{D(P)\}$, \mathbf{M} a model of SF and I a minimal witness for $D(P)$.*

(1) *If $CIR_p(T)$ proves*

$$(\forall \underline{x} \in {}_dP)(\exists \underline{z} \in {}_dP)A[\underline{x}, \underline{z}],$$

then there exists an $\alpha < \varepsilon_0$ such that for all $n < \omega$

$$\mathbf{M} \models (\forall \underline{x} \in {}_dI_{<n})(\exists \underline{z} \in {}_dI_{<f_\alpha(n)})A[\underline{x}, \underline{z}].$$

(2) *Now assume that, in addition, SF is Peano arithmetic and \mathbf{M} the standard structure of the natural numbers. Then we have for all constants \underline{a} :*

$$CIR_p(T) + (IND_N) \vdash \underline{a} \in {}_dP \Rightarrow \underline{a} \in {}_dI_{<\eta_0}.$$

3. Final remarks

The previous sections were devoted to two conceptually fairly different approaches to non-monotonic reasoning by axiomatic extensions. The

closed-world assumption extends the given theory by negations of some non-provable ground atoms whereas (predicate) circumscription adds minimality conditions in the form of induction principles. We end this paper with some general remarks concerning possible extensions of these notions and the relationship between them.

One of the stumbling blocks for the wide use of the closed-world assumption is its inconsistency in the case of indefinite data (cf. REITER 1978). MINKER (1982) overcomes these problems by introducing a modification of the CWA which he has baptized *generalized closed-world assumption* (GCWA). His ideas are extended for example by NAQVI (1986).

A further generalization of the closed-world assumption is the following: Let Γ be a collection of L sentences and T an L theory. The closed world of T with respect to Γ is then defined to be the theory

$$CWA_{\Gamma}(T) := T + \{\neg A : A \in \Gamma \ \& \ T \not\vdash A\}.$$

This version of the closed-world assumption was mentioned by JÄGER (1987) and is a natural extension of the relativized closed-world assumption studied in Section 1. The proper choice of the sets Γ is crucial for all applications but not much is known about the logical and mathematical properties of this concept. It would be interesting to see whether similar results as Theorem 1.8 and Theorem 1.11 can be achieved.

Various extensions and modifications of predicate circumscription are mentioned by McCARTHY (1984) and LIFSCHITZ (1986). They refer to the introduction of *priorities* and *pointwise* circumscription. JÄGER (1986) proves that an analogue to Corollary 2.7 holds for various forms of prioritized circumscription.

Following the techniques of iterated inductive definitions (see e.g. BUCHHOLZ *et al.* 1981), we introduce (JÄGER 1986) the concept of iterated positive disjunctive circumscription and state some proof-theoretic results. There are also close connections between iterated circumscription and the stratified programs of APT *et al.* (1988).

Our personal approach to circumscription is motivated by the theories for inductive definitions and has a proof-theoretic and recursion-theoretic flavour. New model-theoretic aspects of circumscription are studied by SCHLIPF (1987).

LIFSCHITZ (1985) and ETHERINGTON *et al.* (1985) are interested in the relationship between closed-world reasoning and circumscription. Their results, however, apply to very special cases only and more general

situations are not yet completely understood. SHEPHERDSON (1984) studies the connections between the closed-world assumption and Clark's predicate completion (CLARK 1978). This is an important concept in logic programming which often corresponds to fixed point theories as they are studied in proof theory (cf. FEFERMAN 1982).

References

- APT, K.R., BLAIR, H. and WALKER, A., 1988, *Towards a theory of declarative knowledge*, in: J. Minker, ed., *Deductive Databases and Logic Programming* (Morgan Kaufmann, Los Altos).
- ARISTOTELES, transl. 1921, *Analytica priora*, translated by Eugen Rolfes, Band 10, Philosophische Bibliothek (Leipzig).
- ARISTOTELES, transl. 1921, *Analytica posteriora*, translated by Eugen Rolfes, Band 11, Philosophische Bibliothek (Leipzig).
- BACON, F., 1830, *Neues Organ der Wissenschaften* (Wissenschaftliche Buchgesellschaft, Darmstadt, 1962), erstmals (Leipzig).
- BACON, F., 1783, *Ueber die Würde und den Fortgang der Wissenschaften* (Wissenschaftliche Buchgesellschaft, Darmstadt, 1966), erstmals (Pest).
- BARWISE, J., 1975, *Admissible Sets and Structures* (Springer, Berlin, Heidelberg, New York).
- BUCHHOLZ, W., FEFERMAN, S., POHLERS, W. and SIEG, W., 1981, *Iterated inductive definitions and subsystems of analysis: recent proof-theoretical studies*, Lecture Notes in Mathematics 897 (Springer, Berlin, Heidelberg, New York).
- CARNAP, R., 1945, *On inductive logic*, Philosophy of Science 12.
- CARNAP, R. and STEGMÜLLER, W., 1958, *Induktive Logik und Wahrscheinlichkeit* (Springer, Wien).
- CHANG, C.C. and KEISLER, H.J., 1973, *Model Theory* (North-Holland, Amsterdam).
- CLARK, K., 1978, *Negation as failure*, in: H. Gallaire and J. Minker, eds., *Logic and Databases* (Plenum Press, New York).
- ESSLER, W.K., 1970, *Induktive Logik: Grundlagen und Voraussetzungen* (Karl Alber, Freiburg, München).
- ETHERINGTON, D., MERCER, R. and REITER, R., 1985, *On the adequacy of predicate circumscription for closed world reasoning*, Computational Intelligence 1.
- FEFERMAN, S., 1982, *Iterated inductive fixed-point theories: application to Hancock's conjecture*, in: Patras Logic Symposium 1980 (North-Holland, Amsterdam).
- GOGUEN, J.A., THATCHER, J.W., WAGNER, E.G. and WRIGHT, J.B., 1977, *Initial algebra semantics and continuous algebras*, J. Assoc. Comput. Mach. 24.
- HUME, D., 1910, *An enquiry concerning human understanding*, in: C.W. Elitot, ed., Harvard Classics, Vol. 37 (New York).
- HUME, D., 1938, *A Treatise on Human Nature, being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, Reprinted with an introduction by J.M. Keynes and P. Straffa (The University Press, Cambridge, England).
- JÄGER, G., 1986, *Some contributions to the logical analysis of circumscription*, in: Proceedings of the 8th International Conference on Automated Deduction, Lecture Notes in Computer Science 230 (Springer, Berlin, Heidelberg, New York, Tokyo).
- JÄGER, G., 1987, *Annotations on the consistency of the closed world assumption*, preprint (Zürich), to appear in J. Logic Programming.
- LIFSCHITZ, V., 1985, *Closed world data bases and circumscription*, Artificial Intelligence 27.

- LIFSCHITZ, V., 1986, *Pointwise circumscription*, preprint (Stanford).
- LLOYD, J.W., 1984, *Foundations of Logic Programming* (Springer, Berlin, Heidelberg, New York, Tokyo).
- LLOYD, J.W. and TOPOR, R.W., 1984, *Making Prolog more expressive*, J. Logic Programming 1.
- LLOYD, J.W. and TOPOR R.W., 1985, *A basis for deductive data base systems*, J. Logic Programming 2.
- LLOYD J.W. AND TOPOR R.W., 1986, *A basis for deductive data base systems II*, J. Logic Programming 3.
- MAHR, B. and MAKOWSKY J.A., 1983, *Characterizing specification languages which admit initial semantics*, in: Proceedings of 8th CAAP, Lecture Notes in Computer Science 159 (Springer, Berlin, Heidelberg, New York, Tokyo).
- MAKOWSKY, J.A., 1985, *Why Horn formulas matter in computer science*, in: Mathematical Foundations of Software Development, Proceedings of the International Joint Conference on Theory and Practice of Software Development (TAPSOFT), Lecture Notes in Computer Science 185 (Springer, Berlin, Heidelberg, New York, Tokyo).
- MALTSEV, A.I., 1971, *Quasi primitive classes of abstract algebras*, in: The Metamathematics of Algebraic Systems, Collected Papers of A.I. Maltsev (North-Holland, Amsterdam).
- MCCARTHY, J., 1980, *Circumscription — a form of non-monotonic reasoning*, Artificial Intelligence 13.
- MCCARTHY, J., 1986, *Applications of circumscription to formalizing common sense knowledge*, Artificial Intelligence 28.
- MCDERMOTT, D., 1982, *Non-monotonic logic II: non-monotonic modal logics*, J. Assoc. Comput. Mach. 29.
- MCDERMOTT, D. and DOYLE, J., 1980, *Non-monotonic logic I*, Artificial Intelligence 13.
- MINKER, J., 1982, *On indefinite data bases and the closed world assumption*, in: Proc. 6th Conf. Automated Deduction, Lecture Notes in Computer Science 138 (Springer-Verlag, Berlin, Heidelberg, New York, Tokyo).
- MOORE, R.C., 1985, *Semantical considerations on nonmonotonic logic*, Artificial Intelligence 25.
- MOSCHOVAKIS, Y.N., 1974, *Elementary Induction on Abstract Structures* (North-Holland, Amsterdam).
- NAQVI, S.A., 1986, *Some extensions to the closed world assumption*, in: International Conference on Database Theory, Lecture Notes in Computer Science 243 (Springer, Berlin, Heidelberg, New York, Tokyo).
- PARIKH, R., 1984, *Logic of knowledge, games and dynamic logic*, Lecture Notes in Computer Science 181 (Springer, Berlin, Heidelberg, New York, Tokyo).
- PHILODEMUS, transl. 1941, *On Methods of Inference — a Study of Ancient Empiricism*, edited and translated by Ph.H de Lacy and E.A. de Lacy (Philadelphia).
- POPPER, K.R., 1935, "Induktionslogik" und "Hypothesenwahrscheinlichkeit", Erkenntnis 5.
- POPPER, K.R., 1971, *Logik der Forschung* (J.C.B. Mohr, Tübingen).
- REITER, R., 1978, *On closed world data bases*, in: H. Gallaire and J. Minker, eds., *Logic and Databases* (Plenum Press, New York).
- REITER, R., 1980a, *A logic for default reasoning*, Artificial Intelligence 13.
- REITER, R., 1980b, *Equality and domain closure in first-order data bases*, J. Assoc. Comput. Mach. 27.
- SCHLIPF, J.S., 1987, *Decidability and definability with circumscription*, Ann. Pure Appl. Logic 35.
- SHEPHERDSON, J.C., 1984, *Negation as failure: a comparison of Clark's completed data base and Reiter's closed world assumption*, J. Logic Programming 1.
- SHEPHERDSON, J.C., 1985, *Negation as failure II*, J. Logic Programming 3.
- SHEPHERDSON, J.C., 1988, *Negation in logic programming*, in: J. Minker, ed., *Deductive Databases and Logic Programming* (Morgan Kaufmann, Los Altos).

INEXACT AND INDUCTIVE REASONING

J. PARIS and A. VENCOVSKÁ

Dept. of Mathematics, Univ. of Manchester, Manchester, M13 9PL, England

Introduction

In attempting to develop expert systems capable of weighing up uncertain and conflicting evidence one is often faced with the following problem: “Given sentences $\theta_1, \dots, \theta_n$ of the propositional calculus and some linear constraints on the weights of belief attached to these sentences what weight of belief should be assigned to a new sentence θ from the language?” The idea here is that the expert has given some weights of belief, or subjective probabilities to certain statements and we wish to know how this effects our belief in some other statement in which we are interested.

Of course in practice the constraints given by the expert are likely to be very simple. However, we shall consider sets S of (linear) constraints of the form

$$\sum_{j=1}^n \alpha_{ij} w(\theta_j) = \beta_i, \quad i = 1, \dots, m$$

where α_{ij}, β_i are real and $w(\theta)$ stands for the belief in θ , $w(\theta) \in [0, 1]$ and $w(\theta) = 1$ means certainty that θ and $w(\theta) = 0$ means certainty that θ does not hold. Furthermore, when talking about a set of constraints we shall assume they are consistent (unless otherwise stated); precisely what is meant by consistent will be defined shortly.

Returning to the above problem a common initial reaction would be to say that the problem is ill-posed since the constraints will not in general determine $w(\theta)$ uniquely. For example, suppose the set of constraints was empty, A is a propositional variable and we ask what should $w(A)$ be. Then in this case it is consistent to take for $w(A)$ any value between 0 and

1. For example we could consistently set $w(A) = \frac{4}{7}$. However, since the set of constraints does not distinguish between A and $\neg A$ and there is no reason to suppose that the property denoted by A could not equally well be denoted by $\neg A$ it would appear that we should also give $\neg A$ weight $\frac{4}{7}$, which appears inconsistent with w being a belief or subjective probability function. Clearly the only value of $w(A)$ which avoids this "inconsistency" is $\frac{1}{2}$.

It appears then that, whilst in the given problem, $w(\theta)$ may not be uniquely determined *per se*, there are other hidden principles of inexact reasoning which may severely limit the possible "consistent" values of $w(\theta)$. In fact as shown by PARIS and VENCOVSKÁ (1988) there are some seven, reasonable, hidden principles which fix $w(\theta)$ uniquely.

In this paper we shall describe these principles and give a characterization of the value $w(\theta)$ which they determine. We then discuss the relevance of this result to practical expert systems. Finally we give an example of the use of these principles in determining $w(\theta)$ in a particularly simple case.

Principles of inexact reasoning

Before we can specify the seven principles, we need to make the original problem more precise. Clearly we are not interested in just this special case of the problem but in the *inference process* N which, given a set S of constraints, picks a weight of belief function $w = N(S)$, $w : \text{Sentences} \rightarrow [0, 1]$, which satisfies the constraints in S and some other required properties of such a function, which in this paper we take to be:

$$\left. \begin{array}{l}
 \text{For } \psi, \varphi \text{ sentences of the propositional calculus,} \\
 \text{(i) if } \vdash(\varphi \leftrightarrow \psi) \text{ then } w(\varphi) = w(\psi) , \\
 \text{(ii) if } \vdash\psi \text{ then } w(\psi) = 1 \text{ and } w(\neg\psi) = 0 , \\
 \text{(iii) } w(\psi \vee \varphi) = w(\psi) + w(\varphi) \text{ for } \psi, \varphi \text{ disjoint (i.e. } \vdash(\psi \rightarrow \neg\varphi) \text{).}
 \end{array} \right\} \quad (1)$$

Note that we can now make precise the meaning of consistency of a set of constraints. Namely a set of constraints is consistent just if there is a function

$$w : \text{Sentences} \rightarrow [0, 1]$$

which satisfies (1) and the constraints.

Our principles of inexact reasoning are based on three desirable properties of N .

(a) Continuity in the parameters in the constraints. We shall take this as principle 0.

(b) For a set of constraints S the value $N(S)$ gives to a sentence θ should not depend on assertions in S which are irrelevant to θ (i.e. N should be consistent in the non-logical sense). For example if A_1, A_2 are propositional variables and S is just $w(A_1) = \frac{1}{3}$ then the value $N(S)$ gives to A_2 should not depend on S at all, i.e. by earlier considerations $N(S)$ should give A_2 weight $\frac{1}{2}$.

(c) For a set of constraints S , $N(S)$ should not make any assumptions beyond those in S (i.e. N should be fair). For example if as above S is just $w(A_1) = \frac{1}{3}$ then $N(S)$ should not give $A_1 \wedge A_2$ weight 0 since this would mean that A_1 and A_2 were judged by the inference process to be disjoint even though S gives no support for this conclusion.

In what follows we shall assume that N is continuous.

Of course (b) and (c) are rather vague. What we shall do now is to expand them into six precise principles.

PRINCIPLE 1. If the constraint sets S_1, S_2 are equivalent on the basis of (1) then $N(S_1) = N(S_2)$; i.e. changing the way the constraints are expressed should not affect the inference process.

PRINCIPLE 2. Let $g: B_1 \cong B_2$ where B_i ($i = 1, 2$) is the Lindenbaum Boolean algebra of equivalence classes $\bar{\theta}$ of sentences θ from a finite language L_i . Suppose S_i is the set

$$\sum_{j=1}^n \alpha_{kj} w(\theta_j^i) = \beta_k, \quad k = 1, \dots, m$$

of constraints from L_i and $g(\bar{\theta}_j^1) = \bar{\theta}_j^2$ for $j = 1, \dots, n$. Then if $g(\bar{\psi}) = \bar{\varphi}$,

$$N(S_1)(\psi) = N(S_2)(\varphi).$$

i.e. if S_1 is a renamed version of S_2 then $N(S_1), N(S_2)$ should agree up to this renaming.

PRINCIPLE 3. If the sets of constraints S_1, S_2 have no propositional variables in common then $N(S_1), N(S_1 + S_2)$ agree on sentences θ from the language of S_1 ; i.e. if S_2 provides no new information about θ the

inference process should disregard S_2 when assigning a weight to θ on the basis of $S_1 + S_2$.

PRINCIPLE 4. If S_1, S_2 are, respectively, the sets of constraints

$$\sum \alpha_{ij} w(\theta_j \wedge \varphi) = \beta_i, \quad w(\varphi) = \gamma, \quad \sum \delta_{kr} w(\psi_r \wedge \neg \varphi) = \nu_k$$

$$\sum \alpha_{ij} w(\theta_j \wedge \varphi) = \beta_i, \quad w(\varphi) = \gamma, \quad \sum \tau_{sq} w(\eta_q \wedge \neg \varphi) = \lambda_s$$

then for any sentence θ

$$N(S_1)(\theta \wedge \varphi) = N(S_2)(\theta \wedge \varphi),$$

i.e. if S_1, S_2 give the same belief in φ and the same beliefs given φ then the inference process should give the same beliefs relative to φ for both sets of constraints. This then is a relativisation principle.

PRINCIPLE 5. If S_1, S_2 are sets of constraints and $N(S_1)$ satisfies S_2 then $N(S_1) = N(S_1 + S_2)$; i.e. if on the basis of S_1 , $N(S_1)$ gives answers which satisfy S_2 , then adding S_2 to S_1 provides no new information (equivalently gives no reason to change beliefs) and should not cause the inference process to alter its assignment.

PRINCIPLE 6. For the particular case of S being

$$w(A_3) = \gamma (\neq 0), \quad w(A_1 \wedge A_3) = \alpha, \quad w(A_2 \wedge A_3) = \beta,$$

$N(S)(A_1 \wedge A_2 \wedge A_3) = \alpha\beta/\gamma$; i.e. relative to A_3 , S gives no dependence between A_1 and A_2 and thus the inference process should treat them as (statistically) independent on A_3 .

Of all the principles only Principle 6 has a statistical rather than logical justification.

The following theorem is proved by PARIS and VENCOVSKÁ (1988). (For a result along similar lines see JOHNSON and SHORE (1980).)

THEOREM. *There is only one inference process N satisfying Principles 1–6 and continuity. This unique inference process is the so-called Maximum Entropy Inference Process (ME).*

In view of its importance we give the following.

A description of the maximum entropy inference process

Given a set S of constraints and a sentence θ let all the propositional variables in S and θ be amongst A_1, \dots, A_n .

Let C_1, \dots, C_{2n} list all sentences of the form

$$A_1^{i_1} \wedge A_2^{i_2} \wedge \dots \wedge A_n^{i_n}, \quad i_i \in \{0, 1\}$$

where $A^0 = A$, $A^1 = \neg A$.

Using the disjunctive normal form theorem and the property of w that

$$w(\psi \wedge \varphi) = w(\psi) + w(\varphi)$$

for disjoint ψ, φ , we can expand S to a system of linear equations

$$B(w(C_1), \dots, w(C_{2n}))^T = (b_1, \dots, b_m)^T.$$

Now let $(\rho_1, \dots, \rho_{2n})$ be the solution to

$$\begin{aligned} B(w_1, \dots, w_{2n})^T &= (b_1, \dots, b_m)^T, \\ \sum w_i &= 1, \quad w_1, \dots, w_{2n} \geq 0, \end{aligned}$$

which maximises $-\sum w_i \log(w_i)$. Then

$$ME(S)(\theta) = \sum_k \rho_{i_k}$$

where the disjunctive normal form for θ (without repeats) is $\bigwedge_k C_{i_k}$.

This is a good definition in that it is independent of the n chosen (subject only to the given constraints and θ).

Practical consequences

From the above theorem it follows that if we accept Principles 0–6 then ME provides the only mode of inexact reasoning. Unfortunately implementation in a practical expert system appears to provide serious problems as the following theorem shows.

THEOREM. *Fix $\varepsilon > 0$. Then given a (consistent) set of constraints S and sentence θ the problem of finding v such that*

$$|ME(S)(\theta) - v| \leq \frac{1}{2} - \epsilon$$

is NP-hard. More precisely given a function F which picks such a v , $\{x \mid F(x) > \frac{1}{2}\}$ is NP-hard.

A proof of this theorem is given by PARIS and VENCOVSKÁ (1988). Thus if we adopt the conventional belief that NP-hard problems are unfeasible then the best we can do in general to find $ME(S)(\theta)$ is to guess the answer $\frac{1}{2}$!

The options then with an expert system which manipulates uncertainties are either to drop some of the principles or to accept the principles and risk occasionally getting a hopelessly incorrect answer. We believe that whilst some of the principles might well be challenged it would be difficult in practice to live with an expert system which openly flaunted them. Thus the more reasonable option seems to us to accept the possibility of the occasional gross error. Of course in practice one may be able to limit oneself to applications where the computation was feasible. feasible.

The uniqueness of the ME inference process also casts a new light on existing expert systems which manipulate uncertainty. For accepting Principles 0–6, the main theorem says that there is a *correct answer* and the problem becomes finding it. However, most contemporary expert systems first find an answer and then “define” it to be “correct”.

An example

We now give a simple example to show how the principles can be used to determine the weight of a sentence subject to a set of constraints.

Suppose the set of constraints S consists of

$$w(A) = w(B) + \frac{1}{2}, \quad w(A \vee B) = 1,$$

where A and B are propositional variables and we wish to determine the weight of A using the principles. For simplicity of notation let $u = N(S)$ where N is the unique inference process satisfying the principles.

Thus, we already know that u satisfies the above two equations and hence by Principle 1, u (equivalently) satisfies

$$2w(A \wedge \neg B) + w(A \wedge B) = \frac{3}{2}, \quad (2)$$

$$w(A \wedge \neg B) + w(A \wedge B) + w(\neg A \wedge B) = 1, \quad (3)$$

$$w(\neg A \wedge \neg B) = 0, \quad (4)$$

Now let S_1 be the set of constraints

$$w(C) + w(\neg C) = 1$$

where C is a new propositional variable. Since u clearly satisfies S_1 , by Principle 5, $u = N(S + S_1) = N(S_2)$ by Principle 1 where S_2 is

$$\begin{aligned} 2w(A \wedge \neg B \wedge C) + 2w(A \wedge \neg B \wedge \neg C) + w(A \wedge B \wedge C) \\ + w(A \wedge B \wedge \neg C) = \frac{3}{2}, \end{aligned} \quad (5)$$

$$\begin{aligned} w(A \wedge \neg B \wedge C) + w(A \wedge \neg B \wedge \neg C) + w(A \wedge B \wedge C) \\ + w(A \wedge B \wedge \neg C) + w(\neg A \wedge B \wedge C) \\ + w(\neg A \wedge B \wedge \neg C) = 1, \end{aligned} \quad (6)$$

$$w(\neg A \wedge \neg B \wedge C) = w(\neg A \wedge \neg B \wedge \neg C) = 0. \quad (7)$$

Now by considering the Boolean algebra isomorphism which transposes the equivalence classes of $A \wedge B \wedge C$ and $A \wedge B \wedge \neg C$ (but does not change S_2) we see by Principle 2 that

$$u(A \wedge B \wedge C) = u(A \wedge B \wedge \neg C) = \frac{1}{2}u(A \wedge B). \quad (8)$$

Hence, by Principle 5,

$$u = N(S_2) = N(S_2 + (8)) = N(S_3)$$

by Principle 1 where S_3 is

$$w(A \wedge \neg B \wedge C) + w(A \wedge \neg B \wedge \neg C) + w(A \wedge B \wedge C) = \frac{3}{4}, \quad (9)$$

$$w(A \wedge \neg B \wedge C) + w(A \wedge \neg B \wedge \neg C) + w(A \wedge B \wedge \neg C) = \frac{3}{4}, \quad (10)$$

$$\begin{aligned} w(A \wedge \neg B \wedge C) + w(A \wedge \neg B \wedge \neg C) + w(A \wedge B \wedge C) \\ + w(A \wedge B \wedge \neg C) + w(\neg A \wedge B \wedge C) + w(\neg A \wedge B \wedge \neg C) = 1, \end{aligned} \quad (6)$$

$$w(\neg A \wedge \neg B \wedge C) = w(\neg A \wedge \neg B \wedge \neg C) = 0. \quad (7)$$

Furthermore, by Principle 5 again $u = N(S_4)$ where S_4 is S_3 together with

$$w(A \wedge \neg B \wedge \neg C) = u(A \wedge \neg B \wedge \neg C), \quad (11)$$

$$w(\neg A \wedge B \wedge \neg C) = u(\neg A \wedge B \wedge \neg C). \quad (12)$$

Again by Principle 1, $u = N(S_5)$ where S_4 is equivalent to S_5 and S_5 is

$$w(A \wedge \neg B \wedge C) + w(A \wedge B \wedge C) = \frac{3}{4} - u(A \wedge \neg B \wedge \neg C), \quad (13)$$

$$w(A \wedge \neg B \wedge C) + w(A \wedge B \wedge \neg C) = \frac{3}{4} - u(A \wedge \neg B \wedge \neg C), \quad (14)$$

$$\begin{aligned} w((A \wedge \neg B \wedge C) \vee (A \wedge B \wedge C) \vee (A \wedge B \wedge \neg C) \vee (\neg A \wedge B \wedge C)) \\ = 1 - u(A \wedge \neg B \wedge \neg C) - u(\neg A \wedge B \wedge \neg C), \end{aligned} \quad (15)$$

$$w(A \wedge \neg B \wedge \neg C) = u(A \wedge \neg B \wedge \neg C), \quad (11)$$

$$w(\neg A \wedge B \wedge \neg C) = u(\neg A \wedge B \wedge \neg C), \quad (12)$$

$$w(\neg A \wedge \neg B \wedge C) = w(\neg A \wedge \neg B \wedge \neg C) = 0. \quad (7)$$

But now by Principle 4, with φ being

$$(A \wedge \neg B \wedge C) \vee (A \wedge B \wedge C) \vee (A \wedge B \wedge \neg C) \vee (\neg A \wedge B \wedge C),$$

$$u(A \wedge \neg B \wedge C) = t(A \wedge \neg B \wedge C),$$

$$u(A \wedge B \wedge C) = t(A \wedge B \wedge C),$$

$$u(A \wedge B \wedge \neg C) = t(A \wedge B \wedge \neg C),$$

$$u(\neg A \wedge B \wedge C) = t(\neg A \wedge B \wedge C)$$

where $t = N(S_6)$ and S_6 is the set of constraints (13), (14), (15). Also by Principle 2, using the isomorphism which sends the equivalence classes of $A \wedge \neg B \wedge C$, $A \wedge B \wedge C$, $A \wedge B \wedge \neg C$, $\neg A \wedge B \wedge C$ to $A \wedge B \wedge D$, $A \wedge \neg B \wedge D$, $\neg A \wedge B \wedge D$, $\neg A \wedge \neg B \wedge D$ respectively, we see that

$$t(A \wedge \neg B \wedge C) = s(A \wedge B \wedge D),$$

$$t(A \wedge B \wedge C) = s(A \wedge \neg B \wedge D),$$

$$t(A \wedge B \wedge \neg C) = s(\neg A \wedge B \wedge D),$$

$$t(\neg A \wedge B \wedge C) = s(\neg A \wedge \neg B \wedge D)$$

where $s = N(S_7)$ and S_7 is

$$\begin{aligned} w(A \wedge B \wedge D) + w(A \wedge \neg B \wedge D) &= \frac{3}{4} - u(A \wedge \neg B \wedge \neg C), \\ w(A \wedge B \wedge D) + w(\neg A \wedge B \wedge D) &= \frac{3}{4} - u(A \wedge \neg B \wedge \neg C), \\ w((A \wedge B \wedge D) \vee (A \wedge \neg B \wedge D) \vee (\neg A \wedge B \wedge D) \vee (\neg A \wedge \neg B \wedge D)) \\ &= 1 - u(A \wedge \neg B \wedge \neg C) - u(\neg A \wedge B \wedge \neg C), \end{aligned}$$

or equivalently

$$\begin{aligned} w(A \wedge D) &= \frac{3}{4} - u(A \wedge \neg B \wedge \neg C), \\ w(B \wedge D) &= \frac{3}{4} - u(A \wedge \neg B \wedge \neg C), \\ w(D) &= 1 - u(A \wedge \neg B \wedge \neg C) - u(\neg A \wedge B \wedge \neg C). \end{aligned}$$

But now Principle 6 immediately gives

$$s(D) \cdot s(D \wedge A \wedge B) = s(A \wedge D) \cdot s(B \wedge D)$$

and hence

$$\begin{aligned} (u(A \wedge \neg B \wedge C) + u(A \wedge B \wedge C) + u(A \wedge B \wedge \neg C) + u(\neg A \wedge B \wedge C)) \\ \cdot u(A \wedge \neg B \wedge C) &= (u(A \wedge \neg B \wedge C) + u(A \wedge B \wedge C)) \\ \cdot (u(A \wedge B \wedge \neg C) + u(\neg A \wedge B \wedge C)). \end{aligned} \quad (16)$$

Now just as we obtained (8) we can show that

$$u(A \wedge \neg B \wedge C) = u(A \wedge \neg B \wedge \neg C) = \frac{1}{2}u(A \wedge \neg B) \quad (17)$$

$$u(\neg A \wedge B \wedge C) = u(\neg A \wedge B \wedge \neg C) = \frac{1}{2}u(\neg A \wedge B) \quad (18)$$

Finally, we can now solve (2) (3), (4), (8), (16), (17) and (18) to obtain

$$u(A) = \frac{\sqrt{13} + 7}{12}.$$

We remark here that this was actually a rather simple example, most examples appear to be considerably worse if we attempt to solve them directly from the principles. Fortunately, it is not necessary to do so since

by the main theorem the answer agrees with the maximum entropy solution and in simple examples like this the solution can be readily found by elementary calculus.

Comparison with the propositional calculus

Obviously one would expect the *ME* inference process to extend the propositional calculus. The following theorem, which holds no surprises, confirms this.

THEOREM. *Suppose the set of constraints S has the form $w(\theta_i) = 1$, $i = 1, \dots, n$, so S just asserts some certainties. Then $ME(S)(\theta) = \alpha/\beta$ where α is the number of true/false valuations making each of $\theta, \theta_1, \dots, \theta_n$ true (in some suitable finite language) and β the corresponding number for just $\theta_1, \dots, \theta_n$. In particular*

$$ME(S)(\theta) = 1 \Leftrightarrow \theta_1, \dots, \theta_n \vdash \theta .$$

References

- JOHNSON, R. and SHORE, J., 1980, *Axiomatic derivation of the principle of maximum entropy and minimum cross entropy*, IEEE Trans. Inf. Theory 26, pp. 26–37.
- PARIS, J. and VENCOVSKÁ, A., 1988, *On the applicability of maximum entropy to inexact reasoning*, Int. J. Approx. Reason., to appear.
- PARIS, J. and VENCOVSKÁ, A., 1988, *A note on the inevitability of maximum entropy*, Submitted to Int. J. Approx. Reason.

PROBLEMS OF ADMISSIBILITY AND SUBSTITUTION, LOGICAL EQUATIONS AND RESTRICTED THEORIES OF FREE ALGEBRAS

V.V. RYBAKOV

Katedra Algebra, Krasnojarski Gos. University, Krasnojarsk, USSR

The necessity of simplifying derivations in formal systems has led to the study of the class of all rules of inference such that the use of these rules in derivations do not change the set of provable formulas. This class has been called the class of *admissible rules* of inference.

Investigations of the class of admissible rules have for the most part dealt with the intuitionistic propositional calculus H of Heyting. A number of results about admissibility and derivability of rules in H have been obtained in MINTS (1976), TSITKIN (1977, 1979). The problem of finding an algorithm which recognizes the admissibility of rules in H was posed in H. Friedman's survey (FRIEDMAN 1975) as problem 40. A.V. Kuznetsov stated a similar problem: He asked if H has a finite basis for the class of admissible rules. An affirmative answer to Kuznetsov's question would imply a positive solution to Friedman's problem.

The *substitution problem* for propositional logics λ (or the problem of logical equations) consists in the recognition given an arbitrary formula $A(x_i, p_j)$, whether there exist formulas B_i such that $A(B_i, p_j) \in \lambda$. For the modal logic $S4$ and for H the solvability of the substitution problem has been a question of long standing.

In this paper we shall describe the solution of the above stated problems. We adopt an algebraic approach, using properties of free algebras. We shall obtain a solution for H by a reduction to the analogous problems for the systems $S4$ and Grz .

We assume that the reader is familiar with the basic results and notations of first-order theories (COHN 1965) and of Kripke semantics (KRIPKE 1963).

1. Introduction

We start by recalling some terminology and notations. As usual we mean by a *modal logic* (m.l.) a set of modal propositional formulas containing all axioms of the minimal normal modal system K and which is closed with respect to substitution, modus ponens and the normalization rule: $A/\Box A$. Correspondingly, a *superintuitionistic logic* (s.l.) is a set of propositional formulas containing all axioms of Heyting's intuitionistic calculus H and closed with respect to modus ponens and substitution.

We shall use a combination of the algebraic semantics and the relational semantics of Kripke. A *modal algebra* (m.a.) is a Boolean algebra with an additional unary operator satisfying the equations: $\Box 1 = 1$, $\neg \Box(\neg x \vee y) \vee (\neg \Box x \vee \Box y) = 1$. Let $\varphi(p_1, \dots, p_n)$ be a modal formula with propositional letters p_1, \dots, p_n . The formula φ is said to be *valid* (or true) in the m.a. \mathfrak{B} (denoted $\mathfrak{B} \models \varphi$) if for all n -tuples (a_1, \dots, a_n) , each $a_i \in \mathfrak{B}$, the equation $\varphi(a_1, \dots, a_n) = 1$ is true.

A *pseudo-boolean algebra* (p.b.a.) \mathfrak{A} is a distributive lattice with a smallest (0) and a greatest (1) element such that for arbitrary elements $a, b \in \mathfrak{A}$ there exists a relative pseudo-complement $a \supset b$ (i.e. $a \supset b$ is the greatest element $x \in \mathfrak{A}$ such that $a \cap x \leq b$), $a \supset 0$ is called the pseudo-complement of a and is denoted by $\neg a$. The definition of validity of a propositional formula φ in \mathfrak{A} is similar to the definition given above.

If \mathfrak{B} is a m.a. (p.b.a.), then $\lambda(\mathfrak{B}) \Leftrightarrow \{\varphi \mid \mathfrak{B} \models \varphi\}$ is the corresponding m.l. (s.l.). It follows from the completeness theorem (LEMMON 1966, MCKINSEY and TARSKI 1948) (which is based on the construction of Lindenbaum algebras) that for each m.l. (s.l.) λ there exists a m.a. (p.b.a.) \mathfrak{B} such that $\lambda = \lambda(\mathfrak{B})$.

Let λ be a m.l. (s.l.). Then $\text{Var}(\lambda)$ denotes the algebraic variety of these m.a. (p.b.a.): $\{\mathfrak{B} \mid \forall \varphi \in \lambda(\mathfrak{B} \models \varphi)\}$. By the completeness theorem $\varphi \in \lambda \Leftrightarrow \forall \mathfrak{B} \in \text{Var}(\lambda)(\mathfrak{B} \models \varphi)$.

We remind the reader of some definitions and notations concerning the relational Kripke semantics (KRIPKE 1963, LEMMON 1966). A frame $\mathcal{F} = \langle T, R \rangle$ is a pair where T is a nonempty set and R a binary relation on T . Let P be some set of propositional letters. A model $\mathfrak{M} = \langle T, R, V \rangle$ is a 3-tuple where $\langle T, R \rangle$ is a frame and V (valuation) is a function from P into the set of all subsets of the set T .

The validity of modal propositional formulas with respect to elements $x \in T$ is defined by induction on lengths of formulas:

$$\forall p_i \in P (x \Vdash_{\mathfrak{M}} p_i \Leftrightarrow x \in V(p_i))$$

$$\begin{aligned}
 x \Vdash_v (A \wedge B) &\Leftrightarrow (x \Vdash_v A) \& (x \Vdash_v B) \\
 x \Vdash_v (A \vee B) &\Leftrightarrow (x \Vdash_v A) \vee (x \Vdash_v B) \\
 x \Vdash_v (A \rightarrow B) &\Leftrightarrow (x \Vdash_v B) \vee \neg(x \Vdash_v A) \\
 x \Vdash_v \neg A &\Leftrightarrow \neg(x \Vdash_v A) \\
 x \Vdash_v \Box A &\Leftrightarrow \forall y \in T((xRy) \Rightarrow y \Vdash_v A) \\
 x \Vdash_v \Diamond A &\Leftrightarrow \exists y \in T((xRy) \& y \Vdash_v A)
 \end{aligned}$$

A formula φ with propositional letters from P is said to be valid in the model $\mathfrak{M}(\mathfrak{M} \Vdash \varphi)$ iff $\forall x \in T (x \Vdash_v \varphi)$. A formula φ is called valid on the frame $\langle T, R \rangle$ ($\langle T, R \rangle \Vdash \varphi$) iff for all valuations V , $\langle T, R, V \rangle \Vdash \varphi$.

For arbitrary frames $\mathcal{F} = \langle T, R \rangle$ the set $\lambda(\mathcal{F}) = \{\varphi \mid \mathcal{F} \Vdash \varphi\}$ is a modal logic. λ is said to be Kripke complete (FINE 1974) if there exists a frame \mathcal{F} such that $\lambda = \lambda(\mathcal{F})$. There exist modal logics which are incomplete in the sense of Kripke (FINE 1974); however, Kripke semantics and modifications thereof are very convenient technical tools.

Next a few words about first-order semantics in the style of Kripke (THOMASON 1972). Let $\langle W, R \rangle$ be some frame. We assign to this frame a modal algebra $\langle W, R \rangle^+$, which is the Boolean algebra of all subsets of the set W with the operator \Box defined by set w with the operator \sqsubset defined by

$$\forall X \subseteq W \quad \Box X = \{a \mid a \in W, \forall b \in W(aRb \Rightarrow b \in X)\}.$$

Let $X_1, \dots, X_n \in \langle W, R \rangle^+$; by $\langle W, R \rangle^+(X_1, \dots, X_n)$ we denote the subalgebra of $\langle W, R \rangle^+$ generated by the elements X_1, \dots, X_n . Then an arbitrary element of this subalgebra has the form $\varphi(X_1, \dots, X_n)$, where φ is some term. Let us choose the valuation V on p_1, \dots, p_n relative to the frame $\langle W, R \rangle$ such that $V(p_i) = X_i$. If $\varphi(X_1, \dots, X_n)$ is an element of $\langle W, R \rangle^+(X_1, \dots, X_n)$ (where φ is some term), then we mean by $\varphi(p_1, \dots, p_n)$ the formula obtained from φ by substituting letters p_i for X_i and logical connectives for the corresponding algebraic operations.

The following lemma is well known (and proved by induction on the length of φ).

LEMMA 1. For arbitrary $x \in W$,

$$x \in \varphi(X_1, \dots, X_n) \Leftrightarrow x \Vdash_v \varphi(p_1, \dots, p_n)$$

Conversely, to each m.a. \mathfrak{B} we associate a frame \mathfrak{B}^+ in the following way: $\mathfrak{B}^+ \doteq \langle T_{\mathfrak{B}}, R \rangle$, where $T_{\mathfrak{B}}$ is the set of all ultrafilters on \mathfrak{B} , and $\forall \nabla_1, \nabla_2 \in T_{\mathfrak{B}} (\nabla_1, R \nabla_2 \Leftrightarrow (\Box x \in \nabla_1 \Rightarrow x \in \nabla_2))$. According to the Jonsón-Tarski-Stone theorem (JONSSON and TARSKI 1951) the mapping $i: \mathfrak{B} \rightarrow (\mathfrak{B}^+)^+$, where $i(x) \doteq \{\nabla | \nabla \in T_{\mathfrak{B}}, x \in \nabla\}$ is a monomorphism into (if \mathfrak{B} is finite then i is onto).

A m.l. λ is said to have the *finite model property* (fmp) if $\lambda = \bigcap_{i \in I} \lambda(\mathfrak{B}_i)$, where each \mathfrak{B}_i is a finite m.a. LEMMON (1966) has shown that this definition is equivalent to the following: $\lambda = \bigcap_{i \in I} \lambda(\mathcal{F}_i)$, where each \mathcal{F}_i is a finite frame.

We now proceed to discuss rules of inference. Let λ be a m.l. or a s.l. and let A_j, B be formulas in the language of λ . Let p_1, \dots, p_n be all letters occurring in these formulas and x_1, \dots, x_n distinct variables. Expressions of the form

$$A_1(x_1, \dots, x_n), \dots, A_m(x_1, \dots, x_n) / B(x_1, \dots, x_n) \quad (1)$$

are called *rules of inference*. (We consider only rules with a finite number of premisses.) Note that in the Polish mathematical literature (ŁÓŚ 1955, 1958) a more general notion of rule is sometimes used, but by the ŁÓŚ-Susko representation theorem (ŁÓŚ 1958), every standard consequence operation is generated by some countable set of rules of type (1). For this reason we consider only rules of the form (1).

The rule (1) is said to be *admissible* in the logic λ iff for all formulas $B_1, \dots, B_n, A_j(B_1, \dots, B_n) \in \lambda, j = 1, \dots, m$, imply $B(B_1, \dots, B_n) \in \lambda$. The rule (1) is called *derivable* in the logic λ if from A_1, \dots, A_m and the set of theorems of λ one may derive B with the help of modus ponens (and the normalization rule, if λ is a m.l.). It is clear that derivability implies admissibility. HARROP'S rule (1960) $(\neg p \supset (q \vee r)) / (\neg p \supset q) \vee (\neg p \supset r)$ is an example of an admissible rule in H which is not derivable in H .

If in (1) we replace some variables by propositional letters p_j , we have expressions of the form

$$A_1(x_i, p_j), \dots, A_m(x_i, p_j) / B(x_i, p_j).$$

This is called a rule with parameters p_j . This rule will be called *admissible* in λ iff $B(B_i, p_j) \in \lambda$ follow from $A_1(B_i, p_j), \dots, A_m(B_i, p_j) \in \lambda$, for arbitrary formulas B_i .

The *substitution problem* (or problem of logical equations) for the logic

λ consists in recognizing for an arbitrary formula $A(x_i, p_j)$ whether there exist formulas B_i such that $A(B_i, p_j) \in \lambda$. It is clear that such formulas exist iff the rule with parameters $A(x_i, p_j)/(p \leftrightarrow \neg p)$ is not admissible in λ .

There exists an algebraic approach to admissibility and logical equations. Let as above $\text{Var}(\lambda)$ be the variety of algebras corresponding to λ and let $\mathcal{F}_\omega(\lambda)$ be the free algebra of countable rank in $\text{Var}(\lambda)$. Let

$$r = [A_1(x_i, p_j), \dots, A_m(x_i, p_j)/B(x_i, p_j)]$$

be a rule (possibly without parameters). We assign to r the quasi-identity r^* :

$$\left(\bigwedge_{k=1}^m A_k(x_i, p_j) \right) = 1 \Rightarrow B(x_i, p_j) = 1,$$

where x_i are variables and p_j are constants which are interpreted in $\mathcal{F}_\omega(\lambda)$ as free generators. The following proposition belongs to the folklore of our subject and goes back to the methods used in Łós (1955, 1958) to construct logical calculi and consequence operations.

LEMMA 2. (A) *The rule r is admissible in the logic λ iff the quasi-identity r^* is valid in the free algebra $\mathcal{F}_\omega(\lambda)$.*

(B) *There exist formulas B_i such that $A(B_i, p_j) \in \lambda$ iff the equation $A(x_i, p_j) = 1$ (where p_j are free generators) is solvable in the free algebra $\mathcal{F}_\omega(\lambda)$.*

PROOF. Suppose that r is not admissible in λ . Then for some formulas C_i , $A_1(C_i, p_j) \in \lambda, \dots, A_m(C_i, p_j) \in \lambda$ and $B(C_i, p_j) \notin \lambda$. Therefore the identities $A_1(C_i, p_j) = 1, \dots, A_m(C_i, p_j) = 1$ are valid in $\text{Var}(\lambda)$. If we interpret our formulas as elements of the free algebra $\mathcal{F}_\omega(\lambda)$ and regard p_j as free generators of $\mathcal{F}_\omega(\lambda)$, then $B(C_i, p_j) \neq 1$. Hence, the quasi-identity r^* is not valid in $\mathcal{F}_\omega(\lambda)$.

Conversely, assume that r^* is not valid in $\mathcal{F}_\omega(\lambda)$. Then $A_1(C_i, p_j) = 1, \dots, A_m(C_i, p_j) = 1$ and $B(C_i, p_j) \neq 1$ in $\mathcal{F}_\omega(\lambda)$, where C_i are suitable elements from $\mathcal{F}_\omega(\lambda)$. If, as above, we reinterpret $B(C_i, p_j), A_k(C_i, p_j)$ as formulas of the logic λ , then $A_k(C_i, p_j) \in \lambda, B(C_i, p_j) \notin \lambda$. Thus r is not admissible in λ .

Let $A(B_i, p_j) \in \lambda$. Then the identity $A(B_i, p_j) = 1$ is true in $\text{Var}(\lambda)$;

since $\mathcal{F}_\omega(\lambda) \in \text{Var}(\lambda)$, B_i are the solutions to $A(x_i, p_j) = 1$ in $\mathcal{F}_\omega(\lambda)$. Conversely, let $A(x_i, p_j) = 1$ admit a solution in $\mathcal{F}_\omega(\lambda)$. Then $A(C_i, p_j) = 1$ for some $C_i \in \mathcal{F}_\omega(\lambda)$. Because $\mathcal{F}_\omega(\lambda)$ is free in $\text{Var}(\lambda)$, we have that $A(C_i, p_j) = 1$ is true in $\text{Var}(\lambda)$ and, therefore, $A(C_i, p_j) \in \lambda$. The lemma is proved.

Note that to every quasi-identity

$$q: \bigwedge_{k=1}^m (g_k(x_i, p_j) = \varphi_k(x_i, p_j)) \Rightarrow f(x_i, p_j) = g(x_i, p_j)$$

there corresponds a rule $q^*: \bigwedge_{k=1}^m (g_k \leftrightarrow \varphi_k) / f \leftrightarrow g$. It is easy to see that q is valid in $\mathcal{F}_\omega(\lambda)$ iff the rule q^* is admissible in λ . Thus the admissibility problem and substitution problem for the logic λ are reduced (by lemma 2) to questions concerning the universal theory of free algebras $\mathcal{F}_\omega(\lambda)$, with the signature extended by adding constants for free generators.

The rule r is said to be a corollary of the rules r_1, \dots, r_n in the logic λ (in symbols $r_1, \dots, r_n \vdash_\lambda r$) iff the consequence of r is derivable from the premisses of r with the help of theorems of λ , the rules r_1, \dots, r_n and modus ponens (and the normalization rule if λ is a m.l.). A set $B(\lambda)$ of admissible rules in the logic λ is called a basis for the admissible rules of λ if each admissible rule in λ is the corollary of rules r_1, \dots, r_n in $B(\lambda)$.

LEMMA 3. (A) *Let r_1, \dots, r_n, r be rules in λ , then $r_1, \dots, r_n \vdash_\lambda r$ iff the quasi-identity r^* is a corollary of the quasi-identities r_1^*, \dots, r_n^* in the variety $\text{Var}(\lambda)$.*

(B) *The set B is a basis for the admissible rules of the logic λ iff $\{r^* | r \in B\} \cup \{A = 1 | A \in \lambda\}$ is a basis for the quasi-identities of the free algebra $\mathcal{F}_\omega(\lambda)$.*

Proofs are given in TSITKIN (1977) and SELMAN (1972). The first part of lemma 3 is obtained in TSITKIN (1977) for the case $\lambda = H$ and immediately transferred to all m.l. and s.l. It is also a corollary of results in SELMAN (1972). The second part of the lemma follows from the first part.

It follows from lemma 3 that the question of a finite basis for the admissible rules of a logic λ reduces to the question whether the free algebra $\mathcal{F}_\omega(\lambda)$ has a finite basis for the set of quasi-identities.

The Gödel translation T establishes a connection between the admissible rules of s.l. and m.l. We recall that the Gödel translation T from

propositional formulas to modal propositional formulas is defined by induction:

$$\begin{aligned} T(p_i) &= \Box p_i, \\ T(A \wedge B) &= T(A) \wedge T(B) \\ T(A \vee B) &= T(A) \vee T(B) \\ T(\neg A) &= \Box \neg \Box T(A) \\ T(A \supset B) &= \Box(T(A) \rightarrow T(B)) \end{aligned}$$

Let λ be a s.l.; the modal associate of λ is the m.l. λ_1 , such that $\forall A(A \in \lambda \Leftrightarrow T(A) \in \lambda_1)$. By the Dummett–Lemmon theorem (DUMMETT and LEMMON 1959) we have for any s.l. $H + X$ (where H is the Heyting calculus, X is some set of formulas and $H + X$ is the smallest s.l. containing X),

$$A \in H + X \Leftrightarrow T(A) \in S4 + T(X)$$

Thus $S4 + T(X)$ is the smallest modal associate of the s.l. $H + X$ (in the class of all extensions of the m.l. $S4$ of Lewis, where $S4 = K + (\Box p \rightarrow p) \wedge (\Box p \rightarrow \Box \Box p)$). In MAKSIMOVA and RYBAKOV (1974) it is shown that for arbitrary s.l. λ there exists a greatest modal associate $\sigma(\lambda)$ (among extensions of $S4$). It is well known that $\sigma(H) = Grz$, where $Grz = S4 + \Box(\Box(p \rightarrow \Box p) \rightarrow p) \rightarrow p$, e.g. it follows from the finite model property (fmp) of Grz (SEGERBERG 1971) and the fact that σ commutes on arbitrary intersections of logics (MAKSIMOVA and RYBAKOV 1974). We also note the following general fact.

PROPOSITION 4. [RYBAKOV, 1981a, 1986g]. *The rule $A_1, \dots, A_n/B$ is admissible in the s.l. λ iff the rule $T(A_1), \dots, T(A_n)/T(B)$ is admissible in the greatest modal associate $\sigma(\lambda)$.*

If \mathfrak{B} is a m.a. or a p.b.a., then the equalities $a_1 = 1, \dots, a_n = 1$ in \mathfrak{B} are equivalent to the single equality $\bigwedge_{k=1}^n a_k = 1$, and the equality $a = b$ is equivalent to the equality $a \leftrightarrow b = 1$. If λ is a m.l. or a s.l. the assertions $A_1 \in \lambda, \dots, A_n \in \lambda$ are equivalent to $(\bigwedge_{k=1}^n A_k) \in \lambda$, and $A \in \lambda$ iff $(A \leftrightarrow 1) \in \lambda$. For this reason we may only consider rules with one premiss, i.e. rules of the form A/B and quasi-identities only of the form $f = 1 \Rightarrow g = 1$.

The problem of finding an algorithm which recognizes the admissibility of rules in H was posed by H. FRIEDMAN (1975, problem 40). A.V. Kuznetsov stated a similar problem: Does H have a finite basis for the class of admissible rules? It is clear that an affirmative solution of Kuznetsov's problem implies a positive solution of Friedman's problem. In order to solve these questions we first consider the analogous problems for the modal systems $S4$ and Grz . Our aim is to obtain in a uniform way the results for the m.l. $S4$ and Grz and the s.l. H . Because of lemma 2 these questions are connected with properties of the modal algebra $\mathcal{F}_\omega(S4)$; thus we are led to a more detailed investigation of its structure.

2. Description of the structure of $\mathcal{F}_n(S4)$

Let λ be a m.l.; the following is the well-known method (RYBAKOV 1981a, SHEHTMAN 1978) to describe the free algebra in $\text{Var}(\lambda)$ by means of models. Let $P_n = \{p_1, \dots, p_n\}$ be a set of propositional letters. Let $\mathcal{T} = \langle W, R, V \rangle$ be a model where $V: P_n \rightarrow 2^W$. The model \mathcal{T} is said to be n -characteristic for the m.l. λ if for every formula A with propositional letters from P_n ,

$$A \in \lambda \Leftrightarrow \langle W, R, V \rangle \Vdash A.$$

From lemma 1 the next lemma immediately follows.

LEMMA 5. *Let the model $\langle W, R, V \rangle$ be n -characteristic for λ . Then the free modal algebra $\mathcal{F}_n(\lambda)$ is isomorphic to the subalgebra $\langle W, R \rangle^+(V(p_i))$ of the modal algebra $\langle W, R \rangle^+$, generated by the elements $V(p_i)$, $1 \leq i \leq n$ as free generators.*

Thus the description of $\mathcal{F}_n(\lambda)$ depends on the choice of an n -characteristic model, and through the same correspondence determines such a model.

Next we introduce some additional notation and definitions. Let $\mathcal{T} = \langle W, R, V \rangle$ be a model and suppose $X \subseteq W$. By $\langle X \rangle$ we denote the set $\{b \mid \exists a \in X (aRb)\}$. If $X = \{a\}$, we write $\langle a \rangle$ instead of $\langle \{a\} \rangle$. Let X be a subset of W such that $\langle X \rangle = X$; then the pair $\langle X, R \rangle$ is called an open subframe of the frame $\langle W, R \rangle$. The 3-tuple $\langle X, R, V_1 \rangle$, where $V_1(p_i) = V(p_i) \cap X$, is called an open submodel of \mathcal{T} . The following is the main property of open submodels: For each formula A on letters from P_n ,

$$\forall a \in X(a \Vdash_{V_1} A \Leftrightarrow a \Vdash_V A).$$

(Here, we refer to validity relative to the submodel and the main model \mathcal{T} respectively.) This property is proved by an easy induction on the length of A .

Henceforth we shall often identify frames and their sets of elements. A subset X of the reflexive, transitive frame $\langle W, R \rangle$ is called a cluster if

$$\exists x \forall y((xRy) \ \& \ (yRx) \Leftrightarrow y \in X).$$

The depth of $a \in W$ is the maximal length of chains of clusters starting with the cluster containing a . By $\mathcal{D}_n(\langle W, R \rangle)$ we denote the set of elements in W of depth no more than n , similarly $\mathcal{L}_n(\langle W, R \rangle)$ is the set of elements in W of depth n . This set we call the n -layer of W .

We now turn to the construction of an n -characteristic model for $S4$. Toward this end, we first construct a sequence of models $\mathcal{T}_k = \langle T_k, R_k, V_k \rangle$, where each \mathcal{T}_k is an open submodel of \mathcal{T}_{k+1} . A set of clusters from an arbitrary quasi-ordered frame is said to be an antichain if every two elements from different clusters are incomparable. By π_i we denote the i th projection function defined on a Cartesian product.

Let $P_n = \{p_1, \dots, p_n\}$ be a set of propositional letters. We introduce the set $T_1 = \bigcup_{X \subseteq Y} O_X$ and the model $\langle T_1, R_1, V_1 \rangle$, where $Y = 2^{P_n}$, $O_X = \{\langle z, X, 1 \rangle \mid z \in X\}$ and R_1 is the unique reflexive and transitive relation such that the O_X form incomparable clusters with respect to R_1 . Let $V_1(p_i) = \{\langle z, X, 1 \rangle \mid p_i \in z\}$.

Suppose that \mathcal{T}_k has been constructed. Let W be the set of antichains of clusters in \mathcal{T}_k containing at least one cluster from $\mathcal{L}_k(\langle T_k, R_k \rangle)$. Let $W_{k+1} = (W \times C) \setminus D$, where $C = \{\langle z, X, k+1 \rangle \mid X \subseteq 2^{P_n}, z \in X\}$ and $D = \{\langle L, \langle z, X, k+1 \rangle \rangle \mid L \in W, L = \{\nabla\}\}$, the cluster with elements $\{\langle z, X, k+1 \rangle \mid z \in X\}$ and valuation $V: \langle z, X, k+1 \rangle \Vdash_V p_i \Leftrightarrow p_i \in z$ is isomorphic to a submodel of the cluster ∇ .} Let $T_{k+1} = T_k \cup W_{k+1}$,

$$\forall x \in W_{k+1} \langle x, y \rangle \in \overline{R_{k+1}} \Leftrightarrow (x = y) \vee$$

$$(\exists z \exists t((z \in \pi_1(x)) \ \& \ (t \in z) \ \& \ (\langle t, y \rangle \in R_k))) \vee$$

$$((y \in W_{k+1}) \ \& \ (\pi_1(x) = \pi_1(y) \ \& \ \pi_2(\pi_2(x)) = \pi_2(\pi_2(y))))).$$

$$R_{k+1} = \overline{R_{k+1}} \cup R_k,$$

$$V_{k+1}(p_i) \doteq V_k(p_i) \cup \{x \mid x \in W_{k+1}, p_i \in \pi_1(\pi_2(x))\},$$

$$\mathcal{T}_{k+1} \doteq \langle T_{k+1}, R_{k+1}, V_{k+1} \rangle.$$

It is easy to see that \mathcal{T}_k is an open submodel of \mathcal{T}_{k+1} and $\mathcal{T}_k = \mathcal{D}_k(\mathcal{T}_{k+1})$. We have the following.

LEMMA 6 (RYBAKOV 1984b). *Let $T_n \doteq \bigcup_{k < \omega} T_k$, $R \doteq \bigcup_{k < \omega} R_k$, $V \doteq \bigcup_{k < \omega} V_k$; then the model $\mathcal{T}(n) \doteq \langle T_n, R, V \rangle$ is n -characteristic for the modal system $S4$.*

The element x of an arbitrary model $\langle W, R, V \rangle$ is said to be *formulistic* iff there exists a formula A such that $\forall y \in W (y \Vdash_V A \Leftrightarrow (x = y))$. The set $x \subseteq W$ is called *formulistic* iff $\forall y \in W (y \Vdash_V A \Leftrightarrow (y \in X))$ for some formula A .

LEMMA 7 (RYBAKOV 1984b). *All elements of the model $\mathcal{T}(n)$ are formulistic.*

From lemma 6 and lemma 5 we obtain

THEOREM 8 (RYBAKOV 1984b). *The free modal algebra of rank n in the variety $\text{Var}(S4)$ is isomorphic to the subalgebra $\langle T_n, R \rangle^+ (V(p_i))$ of the algebra $\langle T_n, R \rangle^+$, generated by the set $\{V(p_i) \mid p_i \in P_n\}$ as free generators.*

The analogous description of the free modal algebra $\mathcal{F}_n(\text{Grz})$ is obtained in RYBAKOV (1986g, pp. 602–605).

3. The criteria for admissibility of rules and the universal theories of free algebras

Let us consider the quasi-identities in the variety $\text{Var}(S4)$, of the signature Σ_f obtained by adding a countable set of constants for free generators in the m.a. $\mathcal{F}_\omega(\lambda)$, where λ is a certain m.l.

Two quasi-identities are said to be equivalent if they are equivalent as universal formulas on the class $\text{Var}(S4)$. We first note that it is sufficient to consider only the quasi-identities of a rather special form. It is convenient to consider as a start the modal operator \diamond (in view of the expressibilities $\diamond x = \neg \square \neg x$ and $\square x = \neg \diamond \neg x$). We also fix the notations $x^0 = x$ and $x^1 = \neg x$.

THEOREM 9. *There is an algorithm to construct, given an arbitrary quasi-identity q of the form $A = 1 \Rightarrow B = 1$, an equivalent quasi-identity $r(q)$ of the form $((\bigvee_j \varphi_j) = 1 \Rightarrow \neg \diamond x_0 = 1)$, where $\varphi_j = \bigwedge_{i=1}^m x_i^{k(j,i,1)} \wedge \bigwedge_{i=0}^m (\diamond x_i)^{k(j,i,2)}$, $k(j, i, 1), k(j, i, 2) \in \{0, 1\}$ and each x_i is either a variable or a constant. Moreover, $r(q)$ and q have the same constants, and all variables of q are variables of $r(q)$. If $r(q)$ is invalid on $\mathfrak{B} \in \text{Var}(S4)$ for a certain assignment of values to the variables x_i , then q is invalid on \mathfrak{B} for the same variable assignment.*

A proof is given in RYBAKOV (1986g).

The quasi-identity $r(q)$ we call the reduced form (r.f.) of q . If a quasi-identity is of the form $r(q)$, we say that it has reduced form or is in reduced form. If q is a quasi-identity, then by $P(q)$ we denote the set of constants with occurrences in q .

Let f be a term in the signature of a m.a. We say that an occurrence of the variable x_i in the term f is strongly modalized if the given occurrence lies within a subterm $\diamond x_i$. We write the term (formula) in the form $f(\diamond x_1, \dots, \diamond x_m, y_1, \dots, y_n)$ to indicate that all the strongly modalized variables of f are among x_1, \dots, x_m . For convenience let now x_i designate variables or constants from Σ_f .

THEOREM 10 (RYBAKOV 1984b). *For any quasi-identity q of the form $f(\diamond x_1, \dots, \diamond x_m, x_{m+1}, \dots, x_n) = 1 \Rightarrow t(\diamond x_1, \dots, \diamond x_m, x_{m+1}, \dots, x_n) = 1$ we may effectively construct an equivalent quasi-identity $r(q)$ of the form $((\bigvee_j \varphi_j) = 1 \Rightarrow \neg \diamond x_0 = 1)$, where $\varphi_j = (\bigwedge_{i=0}^{n+k} (\diamond x_i)^{k(j,i,2)} \wedge (\bigwedge_{i=m+1}^{n+k} x_i^{k(j,i,1)}))$ and the x_i are variables or constants from Σ_f . Furthermore, if \mathfrak{B} is a modal algebra in $\text{Var}(\lambda)$ and $\neg(\mathfrak{B} \models r(q)_{(a_i)}^{x_i})$, then $\neg(\mathfrak{B} \models q_{(a_i)}^{x_i})$.*

The quasi-identity $r(q)$ from theorem 10 we call the strongly reduced form (s.r.f.) of q . If a quasi-identity has the form $r(q)$ we say that it has s.r.f. or is in s.r.f.

Let q be a quasi-identity in the signature Σ_f and let $r(q)$ be the s.r.f. of q . We introduce some additional notation:

$$\begin{aligned} \theta_1(\varphi_j) &= \{x_i \mid (i > m \ \& \ k(j, i, 1) = 0) \\ &\quad \vee (i \leq m \ \& \ k(j, i, 2) = 0)\}, \\ \theta_2(\varphi_j) &= \{x_i \mid k(j, i, 2) = 0\}, \\ \theta_0(r(q)) &= \{\varphi_j \mid k(j, 0, 2) = 0\}. \end{aligned}$$

Let $\mathcal{D}(r(q))$ be the set of all disjuncts φ_i in the premiss of $r(q)$ satisfying the property $\theta_1(\varphi_i) \subseteq \theta_2(\varphi_i)$.

Before we turn to further results we construct a model on a subset of $\mathcal{D}(r(q))$. Let \mathcal{X} be some subset of $\mathcal{D}(r(q))$. We introduce the model $\langle \mathcal{X}, R, V \rangle$, where $\varphi_i R \varphi_j \Leftrightarrow \theta_2(\varphi_i) \supseteq \theta_2(\varphi_j)$ and the valuation V on the set $P(r(q))$ and the set of all variables from $r(q)$ (the members of both these sets we consider as propositional letters) is defined by the equation

$$V(x_i) = \{ \varphi_j | x_i \in \theta_1(\varphi_j) \}.$$

As above, $P(r(q))$ is the set of all constants from $r(q)$. Let $P_1(r(q)) = \{ x_i | i \leq m \} \cap P(r(q))$. We recall that constants from $P(r(q))$ are interpreted on $\mathcal{F}_\omega(\lambda)$, $\lambda \supseteq S4$ as free generators.

THEOREM 11 (RYBAKOV 1986f). (*Criterion of validity.*) *Let q be a quasi-identity in s.r.f. Then q is invalid on $\mathcal{F}_\omega(S4)$ iff there exists some $\mathcal{X} \subseteq \mathcal{D}(q)$ for which $\langle \mathcal{X}, R, V \rangle$ has the following properties:*

- (1) $\mathcal{X} \cap \theta_0(q) \neq \emptyset$.
- (2) $\varphi_i \Vdash_V \varphi_j$ for each element φ_j of the model $\langle \mathcal{X}, R, V \rangle$.
- (3) Suppose κ is an antichain of clusters in the model $\langle \mathcal{X}, R, V \rangle$, and $\bar{\kappa} \neq 1$. (A special case here corresponds to $\kappa = \emptyset$.) Let $X \subseteq 2^{P(q)}$, and suppose furthermore

$$\forall x, y \in X \forall p_i \in P_1(q) (p_i \in x \Leftrightarrow p_i \in y) \quad (2)$$

and

$$\forall p_i \in P_1(q) \forall x \in X \left(p_i \in \bigcup_{\varphi_\alpha \in M \in \kappa} \theta_2(\varphi_\alpha) \Rightarrow p_i \in x \right) \quad (3)$$

Then there exists a subset $\lambda(\kappa)$ of some cluster from \mathcal{X} such that $\{ \theta_1(\varphi_\alpha) \cap P(q) | \varphi_\alpha \in \lambda(\kappa) \} = X$, $\overline{\lambda(\kappa)} = \bar{X}$, and

$$\forall \varphi_\alpha \in \lambda(\kappa) \left[\theta_2(\varphi_\alpha) = \left(\bigcup_{\varphi_\beta \in \lambda(\kappa)} \theta_1(\varphi_\beta) \right) \cup \left(\bigcup_{\varphi_\beta \in M \in \kappa} \theta_2(\varphi_\beta) \right) \right].$$

- (4) If C is a cluster from $\langle \mathcal{X}, R \rangle$ and $X \subseteq 2^{P(q)}$ and X has properties (2), (3) and $X \not\subseteq \{ \theta_1(\varphi_\alpha) \cap P(q) | \varphi_\alpha \in C \}$, then there exists a subset

$\lambda(C)$ of some cluster from $\langle \mathcal{X}, R \rangle$ such that

$$\{\theta_1, (\varphi_\alpha) \cap P(q) \mid \varphi_\alpha \in \lambda(C)\} = X, \overline{\lambda(C)} = \overline{X},$$

$$\forall \varphi_\alpha \in \lambda(C) \left(\theta_2(\varphi_\alpha) = \left(\bigcup_{\varphi_\beta \in \lambda(C)} \theta_1(\varphi_\beta) \right) \cup \theta_2(\varphi_\gamma) \right), \varphi_\gamma \in C.$$

The proof of this theorem (RYBAKOV 1986, pp. 177–194) is complicated and therefore omitted. It uses the description of $\mathcal{F}_n(S4)$ given in theorem 8.

Using theorem 11 we can recognize the validity of quasi-identities in s.r. f. on $\mathcal{F}_\omega(S4)$. Therefore, from theorems 10, 11 and the disjunction property of the modal system S4 one may obtain the next general result.

Let $A(\bar{x})$ be a universal formula in the signature Σ_f . An *obstacle* for $A(\bar{x})$ on the m.a. \mathfrak{B} is any tuple \bar{a} from \mathfrak{B} such that $\neg(\mathfrak{B} \models A(\bar{a}))$.

THEOREM 12 (RYBAKOV 1986f). *The universal theory of the free modal algebra $\mathcal{F}_\omega(S4)$ in the signature Σ_f is solvable. There exists an algorithm for construction of obstacles for the universal formulas of the signature Σ_f that fail in $\mathcal{F}_\omega(S4)$.*

With the help of theorem 11 in RYBAKOV (1986f, pp. 196–203) one obtains

THEOREM 13 (RYBAKOV 1986f). *Let $q = (A(x_i, p_j) = 1 \Rightarrow B(x_i, p_j) = 1)$ be a quasi-identity in the signature Σ_f of $\mathcal{F}_\omega(H)$ extended by a countable set of constants for free generators. Then q is valid on $\mathcal{F}_\omega(H)$ iff*

$$T(q) = ((T(A)(\Box x_i, \Box p_j) = 1) \Rightarrow (T(B)(\Box x_i, \Box p_j) = 1))$$

is valid on $\mathcal{F}_\omega(S4)$.

From theorems 12, 13 and the disjunction property of H we obtain the next theorem.

THEOREM 14 (RYBAKOV 1986f). *The universal theory of the free p.b.a. $\mathcal{F}_\omega(H)$ in the signature Σ_f extended by adding constants for free generators is solvable. There exists an algorithm to construct obstacles for universal formulas in the signature Σ_f that fail on $\mathcal{F}_\omega(H)$.*

We now describe the criterion of validity on $\mathcal{F}_\omega(\text{Grz})$. We agree to follow the notation introduced for quasi-identities in s.r.f. Let q be a quasi-identity in r.f. The definitions of $\theta_1(\varphi_j)$, $\theta_2(\varphi_j)$, $\mathcal{D}(q)$ we introduced above; let $P_1(\varphi_j) = \theta_1(\varphi_j) \cap P(q)$ and let \mathcal{X} be an arbitrary subset of the set $\mathcal{D}(q)$. We define the model $\langle \mathcal{X}, \triangleleft, V \rangle$, where

$$\forall \varphi_i, \varphi_j \in \mathcal{X} (\varphi_i \triangleleft \varphi_j \Leftrightarrow (\varphi_i = \varphi_j) \vee (\theta_2(\varphi_i) \supseteq \theta_2(\varphi_j)))$$

and the valuation V on $P(q)$ and the set of all variables from q is defined by the equation $V(x_i) = \{\varphi_j | x_i \in \theta_1(\varphi_j)\}$ where x_i ranges over both variables and constants.

The reader will easily see that the relation \triangleleft is a partial order, i.e. $\langle \mathcal{X}, \triangleleft \rangle$ is a poset.

For each $\varphi_j \in \mathcal{X}$ let us fix a (possibly empty) subset $T(\varphi_j)$ of the set $\mathcal{D}(q)$ such that $\varphi_j \notin T(\varphi_j)$ and

$$\forall \varphi_k \in T(\varphi_j) (\theta_2(\varphi_k) = \theta_2(\varphi_j)).$$

We consider all the $T(\varphi_j)$ and \mathcal{X} as disjoint sets even if they really have non-empty intersections. The relation \leq on the set $\mathcal{X} \cup (\bigcup_{\varphi_j \in \mathcal{X}} T(\varphi_j))$ is defined as the reflexive and transitive closure of the relation $(\triangleleft) \cup (\leq_1)$, where

$$\forall \varphi_k \in T(\varphi_j) (\varphi_k \leq_1 \varphi_j).$$

It is easy to see that \leq is a partial order. On the frame $\langle \mathcal{X} \cup (\bigcup_{\varphi_j \in \mathcal{X}} T(\varphi_j)), \leq \rangle$ the valuation V is defined as above:

$$V(x_i) = \{\varphi_j | x_i \in \theta_1(\varphi_j)\}.$$

We recall that constants from $P(q)$ are interpreted as free generators on $\mathcal{F}_\omega(\text{Grz})$.

THEOREM 15. *Let q be a quasi-identity in reduced form. Then q is invalid on $\mathcal{F}_\omega(\text{Grz})$ iff there exists a set $\mathcal{X} \subseteq \mathcal{D}(q)$ where for each $\varphi_j \in \mathcal{X}$ there exists a set $T(\varphi_j) \subseteq \mathcal{D}(q)$ such that*

$$\forall \varphi_k \in T(\varphi_j) (\theta_2(\varphi_k) = \theta_2(\varphi_j)), \quad \varphi_j \notin T(\varphi_j)$$

and the model $\langle \mathcal{X} \cup (\bigcup_{\varphi_j \in \mathcal{X}} T(\varphi_j)), \leq, V \rangle$ has the following properties

- (1) *There is a $\varphi_j \in \mathcal{X}$ such that $k(j, 0, 2) = 0$.*
- (2) *$\varphi_j \Vdash_{\nu} \varphi_j$ for each element φ_j of this model.*
- (3) *If κ is a subset of this model and A is a subset of the set $P(q)$, then there exists an element $\varphi(\kappa, A)$ of this model such that $P_1(\varphi(\kappa, A)) = A$ and*

$$\theta_2(\varphi(\kappa, A)) = \theta_1(\varphi(\kappa, A)) \cup \left(\bigcup_{\varphi_\beta \in \kappa} \theta_2(\varphi_\beta) \right)$$

The proof is complicated and omitted. Theorem 15 gives us an algorithm to recognize valid quasi-identities in r.f. on the free algebra $\mathcal{F}_\omega(Grz)$. From theorems 15, 9 and the disjunction property of Grz we may obtain a result similar to theorem 12 for the free modal algebra $\mathcal{F}_\omega(Grz)$. Using lemma 2, theorems 9, 10, 11, 12, 13, 14, 15 we obtain criteria for the admissibility of rules of inference in the modal systems $S4$, Grz and the Heyting calculus H . All these criteria give algorithms for the recognition of admissibility in the logics $S4$, Grz and H , and we have the following:

THEOREM 16. *The problem of admissibility for rules of inference with parameters (and, as a special case, without parameters) is algorithmically decidable for the logics $S4$, Grz and H .*

As a corollary we obtain a possible solution to FRIEDMAN's problem (1975, problem 40) about algorithmical recognition of the rules of inference admissible in H .

The search for algorithms to recognize admissibility in $S4$, Grz and H is now rather complete. In suitable cases it is more convenient to use semantical criteria. For $S4$ this criterion looks as follows:

THEOREM 17. *The rule A/B is admissible in $S4$ iff $\forall n(\langle T_n, R \rangle^+ \models (A = 1 \Rightarrow B = 1))$.*

There exist similar criteria for Grz and H . Using these, one may show that the rule

$$((q \supset r) \supset (\neg r \vee q)) / (\neg \neg q \vee \neg r)$$

is admissible although underivable in the Heyting calculus H . The rule

$$((\neg\neg p \supset p) \supset (p \vee \neg p)) / (\neg\neg p \vee \neg p)$$

of Scott–Jancov–Kuznetsov is a corollary to this rule.

The decidability of the universal theories of the free algebras $\mathcal{F}_\omega(\lambda)$ in the signature Σ_f , for $\lambda = S4, Grz, H$ (theorem 14, theorem 12 and its analogue for *Grz*) together with lemma 2 imply

THEOREM 18 (RYBAKOV 1986f). (1) *There exists an algorithm for checking solvability in free algebras $\mathcal{F}_\omega(\lambda)$ for finite systems of equations and inequations in the signature Σ_f , and for constructing solutions to them in the cases when $\lambda = S4, Grz, H$.*

(2) *The substitution problem is decidable for the logics $S4, Grz$ and H .*

Thus the decidability of universal theories gives us a way to obtain the decidability for problems about logical equations in the logics $S4, Grz, H$.

We remark that, unlike the universal theories, the elementary theories of the algebras $\mathcal{F}_\omega(\lambda)$ for $\lambda = S4, Grz, H$ are hereditarily undecidable (RYBAKOV 1985d), even in signatures without constants.

4. Bases of admissible rules

Consider Kuznetsov's problem about finite bases of admissible rules in the calculus H and the analogous problems for the systems $S4$ and Grz . We obtained a solution for the calculus H by reduction to the modal systems $S4$ or Grz .

Lemma 3 gives us a method to investigate bases of admissible rules in the logic λ through bases of quasi-identities of the free algebra $\mathcal{F}_\omega(\lambda)$.

The general method to prove that the quasi-identities of an algebra have no basis with only a finite number of variables, is to find a sequence of algebras with the following property: Every n -generated subalgebra of the n th member of the sequence is a member of the quasivariety generated by the given algebra, and no member of the sequence is itself included in this quasivariety. We use this well-known method. The problem is to find such a sequence and to prove these properties.

We introduce a sequence E_n^i , $i < \omega$ of posets. Let $E_n^1 \Rightarrow \langle E_n^1, \leq_1 \rangle$ be a $(2^n + 1)$ -element antichain. Let $P(E_n^1) = \{X \mid X \subseteq E_n^1, 2 \leq \bar{X} \leq 2^n\}$ and $E_n^2 = \langle E_n^1 \cup P(E_n^1), \leq_2 \rangle$, where $(\leq_2) = (\leq_1) \cup (\bar{\leq})$,

$$x \bar{\leq} y \Leftrightarrow ((x = y) \vee (y \in E_n^1 \wedge x \in P(E_n^1) \wedge y \in x)).$$

Now suppose that E_n^i has already been constructed, where $i \geq 2$. We then write $P(E_n^i)$ for the set of all nontrivial antichains in E_n^i that contain at least one element of depth i . We set $E_n^{i+1} = \langle E_n^i \cup P(E_n^i), \leq_{i+1} \rangle$ where $(\leq_{i+1}) = (\leq_i) \cup (\bar{\leq}_i)$,

$$x \bar{\leq} y \Leftrightarrow ((x = y) \vee (y \in E_n^i \wedge x \in P(E_n^i) \wedge y \in x)).$$

Let $E_n = \bigcup_{i < \omega} E_n^i$, $(\leq) = \bigcup_{i < \omega} (\leq_i)$.

If we take the algebras E_n^+ associated with the posets E_n , $n < \omega$, we obtain the needed sequence of modal algebras.

LEMMA 19 (RYBAKOV 1985c). *Every n -generated subalgebra of E_n^+ is a member of the quasivariety generated by $\mathcal{F}_\omega(S4)$.*

A proof is given in RYBAKOV (1985c, pp. 93–97).

LEMMA 20 (RYBAKOV 1985c). *For $n \geq 2$ the algebra E_n^+ is not in the quasivariety generated by $\mathcal{F}_\omega(S4)$.*

A proof is given in RYBAKOV (1985c, pp. 97–101).

From lemmas 19 and 20 we now obtain

THEOREM 21 (RYBAKOV 1985c). *The quasivariety generated by $\mathcal{F}_\omega(S4)$ does not have a basis of quasi-identities in finitely many variables.*

From lemmas 20, 19 and theorems 21, 13 we obtain

THEOREM 22 (RYBAKOV 1985c). *The free p.b.a. $\mathcal{F}_\omega(H)$ has no basis of quasi-identities in finitely many variables.*

In RYBAKOV (1985e, pp. 333–336) it is shown that lemmas 19 and 20 still hold if we take *Grz* instead of *S4*. Hence a counterpart to theorem 21 holds for the free algebra $\mathcal{F}_\omega(\text{Grz})$.

From this result and theorems 21, 22 and lemma 3 we obtain

THEOREM 23 (RYBAKOV 1985c,e). *The modal systems *S4*, *Grz* and the intuitionistic propositional calculus *H* have no basis in a finite number of variables (and in particular have no finite basis).*

Thus we have obtained a negative solution to Kuznetsov's problem about finite bases of admissible rules in *H*.

Although we have no finite bases of admissible rules in $S4$, Grz and H , we may point out some infinite effective bases. For example, a basis of admissible rules for $S4$ is formed by the rules corresponding to quasi-identities in s.r.f. that do not satisfy the conclusion of theorem 11. (Here we take quasi-identities without constants.)

We have not been able to conclude in general that every rule admissible in λ_1 is also admissible in λ_2 whenever $\lambda_1 \subseteq \lambda_2$. Indeed, such a general claim is not true. But for the modal systems $S4$ and Grz it holds:

THEOREM 24. *Every rule of inference without parameters admissible in $S4$ is also admissible in Grz .*

5. Some open problems

(1) Is admissibility always decidable (i.e. does there always exist an algorithm for the recognition of admissibility) in the m.l. (s.l.) λ if λ itself is decidable (decidable and finitely axiomatizable)?

(2) Does there exist an algorithm for the recognition of the rules of inference admissible in all m.l. (s.l.)?

(3) Let λ be a s.l. Is the rule A/B admissible in λ iff the rule $T(A)/T(B)$ is admissible in $S4 + T(\lambda)$?

(4) Let λ be a s.l. Is the rule r admissible in $\sigma(\lambda)$ if r is admissible in $S4 + T(\lambda)$? (Cf. theorem 24.)

(5) Does the modal logic $S4 + T(\lambda)(\sigma(\lambda))$ have a finite basis of admissible rules if the s.l. λ has such a basis?

(6) Does every tabular m.l. (s.l.) have a finite basis of admissible rules?

(7) Does there exist an algorithm for the recognition of admissible rules in an arbitrary finitely axiomatizable m.l. (s.l.) λ of finite layer (i.e. $\lambda \supseteq S4 + \sigma_k(\lambda \supseteq H + I_k)$)?

(8) Do the logics $S4$, Grz , H have independent bases of admissible rules? The algebraic equivalent is the following: Do the free algebras $\mathcal{F}_\omega(\lambda)$, for $\lambda = S4$, Grz , H have independent bases of quasi-identities?

References

- COHN, P.M., 1965, *Universal algebra*, New York.
 DUMMETT, M. AND LEMMON, E., 1959, *Modal logics between $S4$ and $S5$* , Z. math. Log. und Grundl. der Math. 5, pp. 250–264.

- FINE, K., 1974, *An incomplete logic containing S4*, Theoria 40 (1), pp. 23–29.
- FRIEDMAN, H., 1975, *One hundred and two problems in mathematical logic*, J. Symb. Logic 40 (3), pp. 113–130.
- HARROP, R., 1960, *Concerning formulas of the types $A \rightarrow B \vee C$, $A \rightarrow \exists xB(x)$ in intuitionistic formal system*, J. Symb. Logic 25, pp. 27–32.
- JONSSON, B. and TARSKI, A., 1951, *Boolean algebras with operators*, Amer. J. Math. 23, pp. 891–939.
- KRIPKE, S., 1963, *Semantical analysis of modal logic I: Normal propositional logic*, Z. math. Log. und Grundle. der Math. 9, pp. 67–96.
- LEMMON, E.J., 1966, *Algebraic semantics for modal logics I, II*, J. Symb. Logic 31, pp. 46–65, 191–218.
- ŁÓŚ, J., 1955, *The algebraic treatment of the methodology of elementary deductive systems*, Studia Logica 2, pp. 151–212.
- ŁÓŚ, J. and SUSZKO, R., 1958, *Remarks on sentential logics*, Indag. Mathematicae 20, pp. 177–183.
- MAKSIMOVA, L.L. and RYBAKOV, V.V., 1974, *Of lattice of normal modal logic*, Algebra i Logika 15, pp. 188–216 (in Russian).
- MCKINSEY, J. and TARSKI, A., 1948, *Some theorems about the sentential calculi of Lewis and Heyting, L*, J. Symb. Logic 13, pp. 1–15.
- MINTS, G.E., 1976, *The derivability of admissible rules*, J. Soviet Math. 6, N.4.
- RYBAKOV, V.V., 1981a, *The admissibility rules of pretabular modal logic*, Algebra i Logika 20, pp. 440–464; English transl. in Algebra and Logic 20 (1981).
- RYBAKOV, V.V., 1984b, *Criterion of admissibility rules in modal system S4 and intuitionistic logic*, Algebra i Logika 23, pp. 546–572; English transl. in Algebra and Logic, 23 (1984).
- RYBAKOV, V.V., 1985c, *The bases of admissible rules of logics S4 and H*, Algebra i Logika 24, pp. 87–107; English transl. in Algebra and Logic 24 (1985).
- RYBAKOV, V.V., 1985d, *Elementary theories of free closure and pseudo-boolean algebras*, Matematicheskie zametki 37 (6), pp. 797–802; English transl. in Mathematical notices USSR, 37 (1985).
- RYBAKOV, V.V., 1985e, *The bases admissible rules of modal system Grz and intuitionistic logic*, Matematicheskii sbornik 128 (3), pp. 321–339; English transl. in Soviet Math. Sbornic.
- RYBAKOV, V.V., 1986f, *The equations in free closure algebras*, Algebra i Logika 25 (2), pp. 172–204; English transl. in Algebra and Logic 25 (1986).
- RYBAKOV, V.V., 1986g, *The decidability by admissibility modal system Grz and intuitionistic logic*, Izvestija Akademii Nauk SSSR, ser. math. 50 (3), pp. 598–619; English transl. in Soviet Izvestia Acad. Sci.
- SEGERBERG, K., 1971, *An essay in classical modal logic*, Vol. 1–3, Filosofiska Studier, Uppsala.
- SELMAN, A., 1972, *Completeness of calculi for axiomatically defined classes of algebras*, Algebra univers. 2 (1), pp. 20–32.
- SHEHTMAN, V.B., 1978, *The Rieger-Nishimura Laddlers*, Dokl. Akad. Nauk SSSR 240 (3), pp. 549–552; English transl. in Soviet Math. Dokl. (1978).
- THOMASON, S.K., 1972, *Semantical analysis of tense logic*, J. Symb. Logic 37, pp. 150–158.
- TSITKIN, A.I., 1977, *On admissible rules of intuitionistic propositional calculus*, Matematicheskii sbornik 102 (2), pp. 314–323; English transl. in Soviet Math. Sbornic.
- TSITKIN, A.I., 1979, *About admissibility rules in intuitionistic calculus*, Semiotika i informatika 12, pp. 59–61 (in Russian).

This Page Intentionally Left Blank

2 Model Theory

This Page Intentionally Left Blank

ON THE EXISTENCE OF END EXTENSIONS OF MODELS OF BOUNDED INDUCTION

A. WILKIE

Mathematical Institute, Oxford Univ., Oxford, England

J. PARIS

Dept. of Mathematics, Univ. of Manchester, Manchester M13 9DL, England

Introduction

An early result by MACDOWELL and SPECKER (1961) is that any model of full Peano Arithmetic, P , has a proper elementary end extension. From results of FRIEDMAN (1971) it follows that any non-standard countable model of Σ_n induction ($I\Sigma_n$) for $n > 1$ is isomorphic to a proper initial segment of itself and hence has a proper end extension to a model of $I\Sigma_n$. This was later extended by Ressayre (see DIMITRACOPOULOS and PARIS 1988) to the case $n = 1$.

For $n = 0$ this result is false since as we note in Proposition 1 if M, K are models of IA_0 and K is a proper end extension of M ($M \subsetneq K$) then M must also satisfy Σ_1 collection ($B\Sigma_1$). However, there are known to be models of IA_0 that do not satisfy $B\Sigma_1$ (see KIRBY and PARIS 1978) and hence do not have proper end extensions to models of IA_0 . This then raises the question of finding necessary and sufficient conditions on a countable model M of IA_0 for M to have a proper end extension to a model of IA_0 . Clearly a necessary condition is that M satisfies $B\Sigma_1$, but is it enough?

In what follows we shall give an affirmative answer to this question under the assumption that M is closed under exponentiation and a negative answer under the hypothesis that the Δ_0 -hierarchy provably collapses in IA_0 . We shall also introduce the notion of a model of IA_0 being IA_0 -full and show that in the presence of $B\Sigma_1$ this condition is sufficient and is implied by all the other currently known natural sufficient conditions.

We note here that if we go below IA_0 to open induction then these difficulties disappear since BOUGHATTAS (1988) has shown, using an extension of the methods of SHEPHERDSON (1965) that any model of open induction has a proper end extension to a model of open induction.

Notation and definitions

Recall that IA_0 is the scheme

$$\forall \vec{x} [\theta(0, \vec{x}) \wedge \forall y (\theta(y, \vec{x}) \rightarrow \theta(y+1, \vec{x}) \rightarrow \forall y \theta(y, \vec{x})]$$

where θ contains only bounded quantifiers (written $\theta \in \Delta_0$) together with a finite set P^- of axioms for the positive part of a commutative discretely ordered ring with identity.

$B\Sigma_1$ is the schema,

$$\forall \vec{x}, y [\forall x < y \exists z \theta(x, y, z, \vec{x}) \rightarrow \exists t \forall x < y \exists z < t \theta(x, y, z, \vec{x})]$$

where $\theta \in \Delta_0$, or, equivalently in IA_0 , $\theta \in \Sigma_1$.

Throughout let M be a countable non-standard model of $IA_0 + B\Sigma_1$.

For Γ a set of sentences in the language of arithmetic M is said to be Γ -full if whenever $\{\theta_i(x_0, \dots, x_i) \mid i \in \mathbb{N}\}$ is a recursive set of Δ_0 formulae and

$$\forall n \in \mathbb{N}, \Gamma \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i)$$

then

$$\forall x_0 \in M \exists x_1 \leq x_0 \forall x_2 \in M \exists x_3 \leq x_2 \dots M \models \bigwedge_{i=0}^{\infty} \theta_i(x_0, \dots, x_i).$$

The abbreviation $IA_0 \vdash \neg \Delta_0 H$ stands for the hypothesis that the Δ_0 hierarchy provably collapses in IA_0 , i.e. there is a fixed n such that for any $\theta \in \Delta_0$ there is an $X \in \Delta_0$ in prenex normal form with at most n alternations of bounded quantifiers such that

$$IA_0 \vdash \theta \leftrightarrow X.$$

Finally it will be useful to have available the pairing function

$$[e, i] = \begin{cases} 2\langle e, i/2 \rangle, & \text{if } i \text{ is even} \\ 2\langle e, (i-1)/2 \rangle + 1, & \text{if } i \text{ is odd} \end{cases}$$

where $\langle x, y \rangle$ is the usual pairing function $\frac{1}{2}(x + y)(x + y + 1) + x$. The reason for preferring $[x, y]$ is that it satisfies

$$[x, 2y + 1] = [x, 2y] + 1.$$

Main results

THEOREM 1. *If $M \subsetneq K \models I\Delta_0$ then $M \models B\Sigma_1$.*

THEOREM 4. *If M is $I\Delta_0$ -full then M has a proper end extension K with $K \models I\Delta_0$.*

THEOREM 5. *Each of the following imply that M is $I\Delta_0$ -full:*

- (1) *M is short Π_1 -recursively saturated;*
- (2) *M is closed under exponentiation;*
- (3) *$I\Delta_0 \vdash \neg \Delta_0 H$ and $\exists \mathbb{N} < \gamma \in M, M \models \forall x \exists y, y = x^\gamma$;*
- (4) *$I\Delta_0 \vdash \neg \Delta_0 H$ and $\exists a \in M \forall b \in M \exists n \in \mathbb{N}, b \leq a^n$;*
- (5) *$I\Delta_0 \vdash \neg \Delta_0 H$ and $\exists M \subsetneq K \models I\Delta_0 + B\Sigma_1$.*

THEOREM 9. *There is a countable model K of $I\Delta_0 + B\Sigma_1$ which is not $I\Delta_0$ -full. Furthermore, assuming $I\Delta_0 \vdash \neg \Delta_0 H$, K does not have a proper end extension to a model of $I\Delta_0$.*

COROLLARY 7. *Assuming $I\Delta_0 \vdash \neg \Delta_0 H$ the following are equivalent:*

- (1) *$\exists M \subsetneq K \models I\Delta_0 + B\Sigma_1$;*
- (2) *M is $(I\Delta_0 + B\Sigma_1)$ -full.*

We conclude this paper with some remarks on the status of the schema $B\Sigma_1$.

Proofs of the theorems

Theorem 1 is very well known but for the sake of completeness we include a proof.

Proof of Theorem 1

Suppose for simplicity

$$M \models \forall x < c \exists y \theta(x, y)$$

with $\theta \in \Delta_0$. Let $b \in K - M$ so $b > M$. Then since $M \subseteq_e K$, K is a Δ_0 elementary extension of M and

$$K \models \forall x < c \exists y < b \theta(x, y).$$

Using IA_0 let b_0 be the least z satisfying

$$K \models \forall x < c \exists y < z \theta(x, y).$$

Then $b_0 - 1$ cannot be in $K - M$ so $b_0 = (b_0 - 1) + 1 \in M$ and

$$M \models \forall x < c \exists y < b_0 \theta(x, y)$$

as required. \square

Before proving Theorem 4 we need two lemmas.

LEMMA 2. *Suppose Γ is a recursive set of sentences in the language of arithmetic. Then there is a single recursive set $\{\theta_i(x_0, \dots, x_i) \mid i \in \mathbb{N}\}$ of Δ_0 formulae such that for all $n \in \mathbb{N}$,*

$$\Gamma \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i)$$

and for any $K \models IA_0 + B\Sigma_1$, K is Γ -full if and only if

$$\forall x_0 \in K \exists x_1 \leq x_0 \forall x_2 \in K \exists x_3 \leq x_2 \dots K \models \bigwedge_{i=0}^{\infty} \theta_i(x_0, \dots, x_i).$$

[We believe this result to be obvious and if the reader is of like mind we suggest that he skips the very messy proof.]

PROOF. Let $W_e, e \in \mathbb{N}$ be a standard enumeration of the recursively enumerable sets with $W_e = \bigcup_n W_e(n)$ where the $W_e(n)$ are a uniformly recursive sequence of finite sets with

$$\emptyset = W_e(0) \subseteq W_e(1) \subseteq W_e(2) \subseteq \dots$$

and

$$|W_e(n+1) - W_e(n)| \leq 1 \quad \text{for } n \in \mathbb{N}.$$

Let $T(n)$ be the set of consequences of Γ provable from Γ with a proof with code at most n .

For $e \in \mathbb{N}$ we define a recursive sequence S_n^e of finite sequences of formulae as follows. Set $S_n^e = \emptyset$ for $n \leq e$. For $n \geq e$ suppose

$$S_n^e = X_0^e(x_0), X_1^e(x_0, x_1), \dots, X_{j_n}^e(x_0, \dots, x_{j_n}) \quad \text{some } j_n \leq n$$

$$W_e(p) = \{X_0^e(x_0), X_1^e(x_0, x_1), \dots, X_{j_n}^e(x_0, \dots, x_{j_n})\} \quad \text{some } p \leq n$$

and, if $S_n^e \neq \emptyset$, $T(n)$ contains

$$\forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^{j_n} X_i^e(x_0, \dots, x_i).$$

To define S_{n+1}^e , suppose that for some $p' \leq n+1$,

$$W_e(p') = W_e(p) \cup \{X_{j_{n+1}}^e(x_0, \dots, x_{j_{n+1}})\} \neq W_e(p) \quad \text{with } X_{j_{n+1}}^e \in \Delta_0$$

and $T(n+1)$ contains

$$\forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^{j_{n+1}} X_i^e(x_0, \dots, x_i).$$

Set

$$S_{n+1}^e = X_0^e(x_0), \dots, X_{j_n}^e(x_0, \dots, x_{j_n}), X_{j_{n+1}}^e(x_0, \dots, x_{j_{n+1}}).$$

Otherwise set $S_{n+1}^e = S_n^e$.

Clearly as e varies, those infinite limit sequences of the S_n run through all recursive sets $\{\varphi_i(x_0, \dots, x_i) \mid i \in \mathbb{N}\}$ of Δ_0 formulae such that for all $n \in \mathbb{N}$,

$$\Gamma \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \varphi_i(x_0, \dots, x_i).$$

Hence if $\psi_n^e(x_0, \dots, x_n)$ is the conjunction of $x_0 = x_0$ and the formulae in S_n^e (if any) then K is Γ -full just if for all e

$$\forall x_0 \in K \exists x_1 \leq x_0 \forall x_2 \in K \exists x_3 \leq x_2 \dots K \models \bigwedge_{i=0}^{\infty} \psi_i^e(x_0, \dots, x_i).$$

Finally then if we set

$$\theta_n(x_0, \dots, x_n) = \psi_i^e(x_{[e,0]}, \dots, x_{[e,i]})$$

where $n = [e, i]$ then

$$\{\theta_n(x_0, \dots, x_n) \mid n \in \mathbb{N}\}$$

is a recursive set of Δ_0 formulae,

$$\Gamma \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i)$$

for all $n \in \mathbb{N}$ and K is Γ -full just if

$$\forall x_0 \in K \exists x_1 \leq x_0 \forall x_2 \in K \exists x_3 \leq x_2 \dots K \models \bigwedge_{i=0}^{\infty} \theta_i(x_0, \dots, x_i). \quad \square$$

LEMMA 3. *Suppose that M is $I\Delta_0$ -full. Then there is a sentence λ such that $M \models \neg\lambda$ and M is $(I\Delta_0 + \lambda)$ -full.*

PROOF. First suppose that M is $(I\Delta_0 + B\Sigma_1)$ -full and let ψ be the sentence

$$\forall x[\forall y, y^x \text{ exists} \rightarrow \forall y, y^{2^x} \text{ exists}].$$

Then

$$I\Delta_0 + B\Sigma_1, \quad I\Delta_0 + B\Sigma_1 + \psi, \quad I\Delta_0 + B\Sigma_1 + \neg\psi$$

all have the same Π_1 consequences. To see this, suppose

$$I\Delta_0 + B\Sigma_1 + \neg\psi \vdash \forall x\theta(x), \quad \theta \in \Delta_0$$

but

$$I\Delta_0 + B\Sigma_1 \vdash \forall x\theta(x),$$

say $K \models I\Delta_0 + B\Sigma_1 + \neg\theta(a)$. By taking an ultrapower of K if necessary we may assume that a^{η^t} exists in K for some $\eta, t > \mathbb{N}$. But then

$$\{x \in K \mid x \leq a^{\eta^n} \text{ some } n \in \mathbb{N}\} \models I\Delta_0 + B\Sigma_1 + \neg\theta(a) + \neg\psi$$

giving the required contradiction. The case for ψ is similar, taking 2 in place of η .

Hence using $B\Sigma_1$ we see that for $\theta_i(x_0, \dots, x_i) \in \Delta_0$, the sentence

$$\forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i)$$

is Π_1 and is provable in $IA_0 + B\Sigma_1 + \neg\psi(IA_0 + B\Sigma_1 + \psi)$ just if it is provable in $IA_0 + B\Sigma_1$. Hence we can take the required λ to be one of ψ or $\neg\psi$.

So now suppose that M is not $(IA_0 + B\Sigma_1)$ -full. If M is $(IA_0 + \neg\eta)$ -full for some $\eta \in B\Sigma_1$ we may take $\lambda = \neg\eta$ so assume that M is not $(IA_0 + \neg\eta)$ -full for any $\eta \in B\Sigma_1$.

Let $\eta_n, 1 \leq n \in \mathbb{N}$ be a recursive enumeration of $B\Sigma_1$ and let

$$IA_0 + \bigwedge_{j=1}^n \eta_j \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i),$$

$$IA_0 + \neg\eta_m \vdash \forall x_0 \exists x_1 \leq x_0 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i^m(x_0, \dots, x_i)$$

where we may assume the θ_i, θ_i^m satisfy Lemma 1 for $IA_0 + B\Sigma_1, IA_0 + \neg\eta_m$, respectively. Clearly we may assume these are uniformly recursive. Then

$$IA_0 \vdash \neg\eta_1 \vee (\neg\eta_2 \wedge \eta_1) \vee \dots \vee \left(\neg\eta_n \wedge \bigwedge_{i=1}^{n-1} \eta_i \right) \vee \bigwedge_{i=1}^n \eta_i$$

So

$$IA_0 \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \psi_i(x_0, \dots, x_i)$$

where $\psi_i(x_0, \dots, x_i)$ is

$$\bigvee_{j=1}^q \left[\bigwedge_{t=0}^{j-1} \theta_t(x_{[0,0]}, x_{[0,1]}, \dots, x_{[0,t]}) \wedge \bigwedge_{s=1}^q \theta_s^j(x_{[j,0]}, x_{[j,1]}, \dots, x_{[j,s]}) \right]$$

and $[q, q] \leq i < [q+1, q+1]$.

Now suppose

$$\forall x_0 \in M \exists x_1 \leq x_0 \forall x_2 \in M \exists x_3 \leq x_2 \dots M \models \psi_i(x_0, \dots, x_i).$$

Then either for some fixed j

$$\forall x_0 \in M \exists x_1 \leq x_0 \forall x_2 \in M \exists x_3 \leq x_2 \dots M \models \bigwedge_{q=0}^{\infty}$$

$$\bigwedge_{s=1}^q \theta_s^j(x_{[j,0]}, x_{[j,1]}, \dots, x_{[j,s]})$$

or for some $i_n \rightarrow \infty$,

$$\forall x_0 \in M \exists x_1 \leq x_0 \forall x_2 \in M \exists x_3 \leq x_2 \dots M \models \bigwedge_{n=0}^{\infty} \bigwedge_{i=0}^{i_n} \theta_i(x_{[0,0]}, x_{[0,1]}, \dots, x_{[0,i]})$$

both of which give a contradiction. The results follows. \square

Proof of Theorem 4

By Lemma 3 pick λ , such that $M \models \neg\lambda$ whilst M is $(I\Delta_0 + \lambda)$ -full. Now introduce into the language new constants \underline{b} for each element b of M and in the obvious way treat M as a structure for this extended language, $LA(M)$ say. M remains $(I\Delta_0 + \lambda)$ -full.

Suppose now that $\Gamma = \Gamma(\underline{c})$ is a finite extension of $I\Delta_0 + \lambda$ in $LA(M)$ and M is Γ -full. Let $\theta(x, \underline{c})$ be a formula of $LA(M)$, $a \in M$. Then either M is $(\Gamma + \neg\exists x \leq \underline{a}\theta(x, \underline{c}))$ -full or M is $(\Gamma + \theta(\underline{b}, \underline{c}))$ -full for some $b \leq a$. For suppose not. Let π be a new constant symbol and by Lemma 2 let

$$\{\theta_n(x_0, \dots, x_n, \pi, \underline{c}) \mid n \in \mathbb{N}\}$$

be a recursive set of Δ_0 formulae such that

$$\Gamma + \theta(\pi, \underline{c}) \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i, \pi, \underline{c})$$

and for any $K \models I\Delta_0 + B\Sigma_1$, K is $(\Gamma + \theta(\pi, \underline{c}))$ -full just if

$$\forall x_0 \in K \exists x_1 \leq x_0 \forall x_2 \in K \exists x_3 \leq x_2 \dots K \vdash \bigwedge_{i=0}^{\infty} \theta_i(x_0, \dots, x_i, \pi, \underline{c})$$

where K is a structure for $LA(M) + \pi$.

Thus

$$\Gamma + \exists x \leq \underline{a}\theta(x, \underline{c}) \vdash \exists x \leq \underline{a} \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i, x, \underline{c})$$

and since M is not $(\Gamma + \theta(\underline{b}, \underline{c}))$ -full for any $b \leq a$,

$$\neg\exists x \leq \underline{a} \forall x_0 \in M \exists x_1 \leq x_0 \forall x_2 \in M \exists x_3 \leq x_2 \dots M \models \bigwedge_{i=0}^{\infty} \theta_i(x_0, \dots, x_i, x, \underline{c}).$$

Replacing $\exists x \leq \underline{a} \forall x_0 \dots$ here by $\forall y$ ($y = \underline{a} \rightarrow \exists x \leq y \forall x_0 \dots$), we see that M is not $(\Gamma + \exists x \leq \underline{a}\theta(x, \underline{c}))$ -full. But then it is easy to see that our assumption that M is not $(\Gamma + \neg\exists x \leq \underline{a}\theta(x, \underline{c}))$ -full either implies that M is not Γ -full, a contradiction as required.

Using this we can now construct a sequence

$$I\Delta_0 + \lambda = \Gamma_0 \subseteq \Gamma_1 \subseteq \Gamma_2 \subseteq \cdots \subseteq \Gamma_n \subseteq \cdots$$

of finite extensions of $I\Delta_0$ such that M is Γ_n -full for each n , $\Gamma_\infty = \bigcup_n \Gamma_n$ is complete for $LA(M)$ and whenever $\exists x \leq \underline{a}\theta(x, \underline{c}) \in \Gamma_\infty$ then for some $b \leq a$, $\theta(\underline{b}, \underline{c}) \in \Gamma_\infty$.

By a standard omitting types argument Γ_∞ has a model K in which the interpretations of the \underline{b} for $b \in M$ form an initial segment. Furthermore since M is Γ_n -full for each n , Γ_∞ contains the Δ_0 theory of M (in $LA(M)$) so M is, up to isomorphism, an initial segment of K . Finally, this initial segment is proper since $K \models \lambda$ whilst $M \models \neg\lambda$. \square

REMARK. Clearly by a similar proof, if M is $(I\Delta_0 + B\Sigma_1)$ -full then M has a proper end extension to a model of $I\Delta_0 + B\Sigma_1$.

Proof of Theorem 5(1)

Suppose that

$$I\Delta_0 \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \cdots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i) \quad (1)$$

where $\{\theta_i(x_0, \dots, x_i) \mid i \in \mathbb{N}\}$ is a recursive set of Δ_0 formulae and that M is short Π_1 -recursively saturated. By (1)

$$\left\{ \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \cdots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i) \mid n \in \mathbb{N} \right\}$$

is finitely satisfiable in M . Let $a_0 \in M$. Then by (1)

$$\left\{ x_1 \leq a_0 \wedge \forall x_2 \exists x_3 \leq x_2 \cdots \bigwedge_{i=0}^n \theta_i(a_0, x_1, \dots, x_i) \mid n \in \mathbb{N} \right\}$$

is recursive, finitely satisfiable in M and, since $M \models B\Sigma_1, \Pi_1$ in M . Hence for some $a_1 \in M$

$$\left\{ \forall x_2 \exists x_3 \leq x_2 \cdots \bigwedge_{i=0}^n \theta_i(a_0, a_1, x_2, \dots, x_i) \mid n \in \mathbb{N} \right\}$$

is finitely satisfiable in M . Clearly continuing in this way shows that

$$\forall x_0 \in M \exists x_1 \leq x_0 \forall x_2 \in M \exists x_3 \leq x_2 \cdots M \models \bigwedge_{i=0}^{\infty} \theta_i(x_0, x_1, \dots, x_i)$$

as required. \square

REMARK. Solovay has shown (see Theorem 4 of PARIS (1981)) that any countable recursively saturated model of $I\Delta_0 + B\Sigma_1$ is isomorphic to a proper initial segment of itself. However, by using the characterization given by DIMITRACOPOULOS and PARIS (1988) of countable models of $I\Delta_0 + B\Sigma_1 + \text{Exponentiation}$ which are isomorphic to proper initial segments of themselves, it can be seen that short Π_1 recursive saturation is not sufficient to ensure isomorphism with a proper initial segment.

Proof of Theorem 5(2)

To prove this theorem we require the following overspill lemma.

LEMMA 6. *Let $K \models I\Delta_0 + B\Sigma_1$ be non-standard and closed under exponentiation and suppose that $\psi(x) \in \Pi_1$ and $I\Delta_0 \vdash \psi(\underline{n})$ for each $n \in \mathbb{N}$. Then $K \models \psi(e)$ for some $e > \mathbb{N}$.*

PROOF. Suppose not. Let $\psi(x) = \forall y \theta(x, y)$ with $\theta \in \Delta_0$. Pick $\mathbb{N} < b \in K$. Then in K for all $c < b$ either $\mathbb{N} < c$, in which case $\exists y \neg \theta(c, y)$ or $c < \mathbb{N}$ in which case $I\Delta_0 \vdash \psi(\underline{c})$ and $M \vdash \exists$ semantic tableaux proof with code $< b$ of $\psi(\underline{c})$ from $I\Delta_0$.

Therefore, since $K \vdash B\Sigma_1$, for some $d \in K$,

$$K \models \forall x < b [\exists y < d \neg \theta(x, y) \vee \exists \text{ sem. tab. proof } < b \text{ of } \psi(\underline{x}) \text{ from } I\Delta_0].$$

Let $a + 1$ be the least $x < b$ such that

$$K \models \exists y < d \neg \theta(x, y).$$

Then $a > \mathbb{N}$, so by assumption $K \models \neg \psi(a)$, whilst $K \models \exists$ sem. tab. proof $< b$ of $\psi(\underline{a})$ from $I\Delta_0$. But this is a contradiction by Theorem 8.10 of PARIS and WILKIE (1987). \square

Returning now to the proof of Theorem 5(2) recall that by a result of LESSAN (1978) (see also Theorem 2 of DIMITRACOPOULOS and PARIS (1982)) there is a $\Gamma(x, y, z) \in \Delta_0$ such that for any $K \models I\Delta_0$ and $\theta(\vec{x}) \in \Delta_0$, $\vec{a} \in K$ if $b \in K$ exceeds $\xi_\theta(\vec{a}) (= 2^{(\max(\vec{a})+2)^{|\theta|}})$ then

$$K \models \theta(\vec{a}) \leftrightarrow \Gamma(b, \langle \vec{a} \rangle, [\theta]). \quad (2)$$

In order to make the following proof more transparent we shall use the more suggestive notation

$$\models_b \theta(\vec{a}) \text{ for } \Gamma(b, \langle \vec{a} \rangle, [\theta]).$$

By (2), for standard $\theta, \psi \in \Delta_0$ and b sufficiently large in relation to parameters, \models_b , provably in $I\Delta_0$, acts like a satisfaction relation. For example in $I\Delta_0$ we can prove for $\varphi(\vec{x}) \in \Delta_0$ and $\psi(\vec{x}) = \exists z \leq x_1 \theta(z, \vec{x}) \in \Delta_0$ that

$$\forall \vec{x} \forall y \geq \xi_\psi(\vec{x}) [\models_y \exists z \leq x_1 \theta(z, \vec{x}) \leftrightarrow \exists z \leq x_1 \models_y \theta(z, \vec{x})]$$

and

$$\forall \vec{x} \forall y_1, y_2 \geq \xi_\varphi(\vec{x}) [\models_{y_1} \varphi(\vec{x}) \leftrightarrow \models_{y_2} \varphi(\vec{x})].$$

Hence by Lemma 6 we can assume that if K is non-standard and closed under exponentiation then any finite number of such properties hold in K for all Δ_0 -formulae (in the sense of K) with codes $\leq e$ for some $e > \mathbb{N}$.

Now fix M to be, as in Theorem 5(2), a non-standard model of $I\Delta_0 + B\Sigma_1$ closed under exponentiation and suppose that

$$\{\theta_n(x_0, \dots, x_n) \mid n \in \mathbb{N}\}$$

is a recursive set of Δ_0 formulae and for each $n \in \mathbb{N}$,

$$I\Delta_0 \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i).$$

We wish to show that

$$\forall x_0 \in M \exists x_1 \leq x_0 \forall x_2 \in M \exists x_3 \leq x_2 \dots M \models \bigwedge_{i=0}^{\infty} \theta_i(x_0, \dots, x_i).$$

Let $\lambda(n)$ be the Π_1 statement

$$\forall m \leq n \forall z \forall y \geq \xi_{\gamma_m}(z)$$

$$\left[\models_y \forall x_0 \leq z \exists x_1 \leq x_0 \forall x_2 \leq z \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^m \theta_i(x_0, \dots, x_i) \right]$$

where $\gamma_m(z)$ is the formula following \models_y . By our assumption (and the provable properties of \models_y) this is provable in $I\Delta_0$ for all $n \in \mathbb{N}$. Hence by Lemma 6 we may assume that for some $e > \mathbb{N}$

$$\forall f \leq e \forall z \forall y \geq \xi_{\gamma_f}(z)$$

$$\left[\models_y \forall x_0 \leq z \exists x_1 \leq x_0 \forall x_2 \leq z \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(x_0, \dots, x_i) \right]$$

holds in M . As remarked earlier, we may further assume that in what follows \models_y acts like a satisfaction relation for Δ_0 -formulae (in the sense of M) provided their codes do not exceed some $c > \mathbb{N}$ and y is sufficiently large in relation to the parameters.

So let $f > \mathbb{N}$ be such that $\lceil \gamma_f \rceil$ is much less than c . Then

$$\forall z \forall y \geq \xi_{\gamma_f}(z) \left[\models_y \forall x_0 \leq z \exists x_1 \leq x_0 \forall x_2 \leq z \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(x_0, \dots, x_i) \right].$$

Let $a_0 \in M$. Then

$$\forall z \geq a_0 \forall y \geq \xi_{\gamma_f}(z) \left[\models_y \exists x_1 \leq a_0 \forall x_2 \leq z \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(a_0, x_1, \dots, x_i) \right]. \quad (3)$$

We now claim that for some $a_1 \leq a_0$

$$\forall z \geq a_0 \forall y \geq \xi_{\gamma_f}(z) \left[\models_y \forall x_2 \leq z \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(a_0, a_1, x_2, \dots, x_i) \right].$$

For suppose not. Then by $B\Sigma_1$ there is $\delta \geq a_0$ such that for all $a_1 \leq a_0 \exists a_0 \leq \delta_1 \leq \delta$ such that

$$\exists y \geq \xi_{\gamma_f}(\delta_1) \left[\neg \models_y \forall x_2 \leq z \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(a_0, a_1, x_2, \dots, x_i) \right].$$

By the assumed overspill of provable properties of \models_y this holds for all $y \geq \xi_{\gamma_f}(\delta)$ and so

$$\forall y \geq \xi_{\gamma_f}(\delta) \left[\forall x_1 \leq a_0 \neg \models_y \forall x_2 \leq \delta \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(a_0, x_1, \dots, x_i) \right]$$

which with overspill properties of \models_y contradicts (3).

Hence, the claim holds and clearly continuing in this way gives the required result. \square

REMARK. A direct proof that any countable model of $IA_0 + B\Sigma_1$ which is closed under exponentiation has a proper end extension to a model of IA_0 may be obtained by mimicking the proof of Theorem 4 but with ‘‘Semantic Tableaux consistency of Γ ’’ in place of ‘‘ Γ -full’’ and adding a new constant symbol $\pi > M$ to ensure that the end extension is proper.

Proof of Theorem 5(3)

The proof of this result has a very similar structure to that of Theorem 5(2). As in that proof, we require an overspill lemma which is of interest in its own right. Before proving it we recall that under the assumption that $I\Delta_0 \vdash \neg \Delta_0 H$, $I\Delta_0$ is finitely axiomatizable.

LEMMA 7. *Let $K \models I\Delta_0 + B\Sigma_1$ and suppose that for some $\aleph < \gamma \in K$, $K \models \forall x \exists y y = x^\gamma$. Suppose further that $I\Delta_0$ is finitely axiomatizable and that $I\Delta_0 \vdash \psi(\underline{n})$ for all $n \in \mathbb{N}$ where $\psi \in \Pi_1$. Then $K \models \psi(\underline{e})$ for some $\aleph < e \in K$.*

PROOF. The proof is similar to that of Lemma 6 but the lack of exponentiation requires us to be a little more careful. As in that proof we find $a \in K$ such that

$$K \models \neg \psi(a) \wedge \exists \text{ sem. tab. proof } p < b \text{ of } \psi(\underline{a}) \text{ from } I\Delta_0$$

where we now use the finite axiomatization of $I\Delta_0$.

We may assume $\aleph < b$ and $2^b < \gamma$.

Thus, p is (the code for) a tree of subsets of

$$\{\varphi_j(t_1, \dots, t_j) \mid 1 \leq j \leq q, t_1, \dots, t_j \text{ terms in } K\}$$

where the $\varphi_j(x_1, \dots, x_j)$ run through the finitely many subformulae of $\neg \psi(x_1) + I\Delta_0$, the root of p is $\neg \psi(\underline{a}) + I\Delta_0$, all tips of p contain some formula and its negation and all vertices are related to their immediate successors by some standard rules for semantic tableaux (see for example Section 8.9 of PARIS and WILKIE (1987)).

Since $K \models \neg \psi(a)$, let c be such that $K \models \neg \theta(a, c)$. Then by induction on the levels of p we can show that there is a vertex V and an assignment $f \in K$ of values to the terms t appearing in the φ_j in V such that $f(t) < \max(a, c, 2)^p$ and under this assignment all the formulae in V are true. [Notice this can be expressed because there are only finitely many φ_j .] In particular this holds at a tip of p giving the required contradiction.

Returning now to the proof of Theorem 5(3) we recall (see DIMITRACOPOULOS and PARIS 1982) that if $I\Delta_0 \vdash \neg \Delta_0 H$ then there is $\Gamma(x, y, z) \in \Delta_0$ such that for any

$$K \models I\Delta_0, \theta(\vec{x}) \in \Delta_0 \text{ and } \vec{a} \in K \text{ if } b \in K, b \geq \xi_\theta(\vec{a}) = \max(\vec{a}, 2)^{n_\theta},$$

where n_θ depends recursively on θ , then

$$K \models \theta(\vec{a}) \leftrightarrow \Gamma(b, \langle \vec{a} \rangle, [\theta]).$$

The proof now proceeds exactly as in Theorem 5(2) with the following minor modification. To use $\models_y \theta(\vec{a})$ for non-standard θ we need $y \geq \xi_\theta(\vec{a})$. In the proof of Theorem 5(2) this causes no problem since it requires only double exponentiation. In the current proof, however, we need $n_\theta \leq \gamma$ to ensure the existence of such y . For this reason we need to further choose f sufficiently small that for all formulae θ under consideration, $n_\theta \leq \gamma$. This is possible by a standard Δ_0 -overspill argument. \square

Proof of Theorem 5(4)

Assume that $I\Delta_0 \vdash \neg \Delta_0 H$ and

$$\forall b \in M \exists n \in \mathbb{N}, \quad b \leq a^n$$

for some $a \in M$. We show that under these assumptions M is short Π_1 -recursively saturated so that the result follows by Theorem 5(1).

Let $\Gamma(x, y, z) \in \Delta_0$ be as in the proof of Theorem 5(3) and suppose that $\{\theta_i(x, \vec{y}) \mid i \in \mathbb{N}\}$ is a recursive set of Π_1 formulae such that

$$x \leq c + \{\theta_i(x, \vec{d}) \mid i \in \mathbb{N}\}$$

is finitely satisfiable in M . Suppose that this set was not satisfiable. Then

$$\forall x \leq c \exists i \in \mathbb{N}, M \models \neg \theta_i(x, \vec{d}).$$

Let $\neg \theta_i(x, \vec{d}) = \exists z \psi_i(z, x, \vec{d})$. Then in M ,

$$\forall x \leq c \exists b, i, j, y, e, f [b = a^i \wedge e = a^j \wedge y \geq \xi_{\psi_i}(f, x, \vec{d}) \wedge \models_y \psi_i(f, x, \vec{d})].$$

By $B\Sigma_1$, the b can be bounded, say by a^m . But this implies that

$$\forall x \leq c \exists i \leq m, M \models \neg \theta_i(x, \vec{d})$$

contradicting the finite satisfiability. \square

Proof of Theorem 5(5)

Assume that $I\Delta_0 \vdash \neg \Delta_0 H$ and $M \underset{c}{\subset} K \models I\Delta_0 + B\Sigma_1$. We may assume that for some $c \in K - M$, c^ν exists for some $\nu > \aleph$. For either this already

holds or by Theorems 5(4) and 4, K has an end extension to a model of $IA_0 + B\Sigma_1$ in which it must hold.

By overspill we may assume that, with the notation of the proof of Theorem 5(3), \models_{c^ν} has some standard properties of a satisfaction relation for $[\theta(\vec{x})] < e$ whenever $\vec{x} \leq c$ and θ is Δ_0 in the sense of K for some $e > \mathbb{N}$.

Now suppose $\{\theta_n(x_0, \dots, x_n) \mid n \in \mathbb{N}\}$ is a recursive set of Δ_0 formulae and for all $n \in \mathbb{N}$,

$$IA_0 \vdash \forall x_0 \exists x_1 \leq x_0 \forall x_2 \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^n \theta_i(x_0, \dots, x_i).$$

By overspill let $f > \mathbb{N}$ be such that

$$\left[\forall x_0 \leq z \exists x_1 \leq x_0 \forall x_2 \leq z \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(x_0, \dots, x_i) \right] < e$$

and

$$\vdash_{c^\nu} \forall x_0 \leq c \exists x_1 \leq x_0 \forall x_2 \leq c \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(x_0, \dots, x_i).$$

Let $a_0 \in M$. Then since $a_0 < c$,

$$\vdash_{c^\nu} \exists x_1 \leq a_0 \forall x_2 \leq c \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(a_0, x_1, \dots, x_i).$$

Hence for some $a_1 \leq a_0$,

$$\vdash_{c^\nu} \forall x_2 \leq c \exists x_3 \leq x_2 \dots \bigwedge_{i=0}^f \theta_i(a_0, a_1, x_2, \dots, x_i).$$

Clearly continuing in this way shows that

$$\forall x_0 \in M \exists x_1 \leq x_0 \forall x_2 \in M \exists x_3 \leq x_2 \dots M \models \bigwedge_{i=0}^{\infty} \theta_i(x_0, \dots, x_i)$$

are required. \square

Corollary 7 follows from the proof of Theorem 5(5) and Theorem 4.

COROLLARY 8. *Suppose that there is no $t \in M$ such that for $\nu \in M$, $2^{[t/\nu]}$ exists in M if and only if $\nu > \mathbb{N}$. Then assuming $IA_0 \vdash \neg \Delta_0 H$ the following are equivalent.*

- (1) $\exists M \subset K \models IA_0$.
- (2) M is IA_0 -full.

PROOF of (1) \Rightarrow (2). Pick $b \in K - M$ and let t be such that $2' \leq b < 2'^{+1}$. If $2^{[t/\nu]} \in K - M$ for some $\nu > \mathbb{N}$ then

$$M \subsetneq \{x \mid \exists n \in \mathbb{N}, x \leq 2^{[n]}\} \subsetneq K$$

and (2) follows by Theorems 4 and 5(5).

Otherwise $2^{[t/\nu]} \in K - M$ just if $\nu \in \mathbb{N}$ so $t \in M$ and the required contradiction follows. \square

REMARK. It would be nice to have the equivalence of (1) and (2) without the assumption on the structure of M (or even better without assuming $I\Delta_0 \vdash \neg \Delta_0 H$). However we are unable to prove this.

Proof of Theorem 9

Let T be a maximal set of Σ_1 sentences such that $T + I\Delta_0$ is consistent. Let K be a full ultrapower of a model of $T + I\Delta_0$ and let J be the initial segment of K in which the Δ_0 -definable elements of K are cofinal. Then $J \subsetneq K$, $J \models I\Delta_0 + B\Sigma_1$ and any Σ_1 sentence true in K is true in J so $J \models T$.

Our initial aim is to construct $M \prec_{\forall \leq \Sigma_1} J$ such that $\Pi_1(M)$ (the Π_1 theory of M) is not coded in M . To this end we construct a sequence of Π_1 formulae $\theta_i(x_0, \dots, x_{n_i})$ satisfiable on J . Suppose that at some stage we have found $\theta_i = \theta_i(x_0, \dots, x_{n_i})$. There are two cases according to the parity of i .

Case 1: i even

Let φ be the $(i/2)$ th formula in some standard enumeration of $\forall \leq \Sigma_1$ formulae. If φ has no initial bounded universal quantifiers, say $\varphi = \exists z_1, \dots, z_q \psi(x_0, \dots, x_m, z_1, \dots, z_q)$ ($m \geq n_i$) with $\psi \in \Delta_0$ put

$$\theta_{i+1} = \theta_i \wedge \psi(x_0, \dots, x_m, x_{m+1}, \dots, x_{m+q})$$

if this is satisfiable in J and put

$$\theta_{i+1} = \theta_i \wedge \neg \varphi \quad \text{otherwise.}$$

If φ has initial bounded universal quantifiers, say

$$\varphi = \forall \vec{z} \leq \vec{x} \exists \vec{y} \psi(x_0, \dots, x_m, z_1, \dots, z_q, \vec{y})$$

($m \geq n_i$) with $\psi \in \Delta_0$ put

$$\theta_{i+1} = \theta_i \wedge \forall \vec{z} \leq \vec{x} \exists \vec{y} \leq x_{m+1} \psi(\vec{x}, \vec{z}, \vec{y})$$

if this is satisfiable in J and

$$\theta_{i+1} = \theta_i \wedge \neg \exists \vec{y} \psi(\vec{x}, x_{m+1}, \dots, x_{m+q}, \vec{y})$$

otherwise. Notice that since $J \models B\Sigma_1$ in each case θ_{i+1} is satisfiable in J .

Case 2: i odd

Let $j = (i - 1)/2$. Since the Δ_0 -definable elements of J are cofinal in J and θ_i is satisfiable there is $\gamma(x) \in \Delta_0$ such that

$$J \models \exists! x \gamma(x) \wedge \exists x \exists x_0, \dots, x_{n_i} \leq x [\theta_i(x_0, \dots, x_{n_i}) \wedge \gamma(x)]. \quad (3)$$

Now let σ be a Π_1 sentence such that

$$I\Delta_0 + B\Sigma_1 \vdash \sigma \leftrightarrow \forall x [\gamma(x) \rightarrow \exists \vec{x} \leq x (\theta_i(\vec{x}) \wedge [\underline{\sigma}] \notin x_j)]$$

where $[\underline{\sigma}] \notin x_j$ is the formula expressing that $[\underline{\sigma}]$ is not a member of the set coded by x_j (in some standard coding).

If $J \models \sigma$ then put

$$\theta_{i+1} = \theta_i \wedge \gamma(x_{n_i+1}) \wedge x_0, \dots, x_{n_i} \leq x_{n_i+1} \wedge [\underline{\sigma}] \notin x_j.$$

Clearly θ_{i+1} is satisfiable in J . On the other hand if $J \models \neg \sigma$ then put

$$\theta_{i+1} = \theta_i \wedge \gamma(x_{n_i+1}) \wedge x_0, \dots, x_{n_i} \leq x_{n_i+1} \wedge [\underline{\sigma}] \in x_j.$$

Again by (3) θ_{i+1} is satisfiable in J .

Having constructed the sequence θ_i let $a_i, i \in \mathbb{N}$ satisfy $\bigwedge_i \theta_i$ in J , this is possible since K is a full ultrapower. Then by considering Case 1 it is easy to check that the set $\{a_i \mid i \in \mathbb{N}\}$ forms the universe of a substructure M of J and that $M < J$. It follows that $M \models I\Delta_0 + B\Sigma_1 + T$. Furthermore the a_i satisfy $\bigwedge_i \theta_i$ in M and hence $\Pi_1(M) (= \Pi_1(J))$ is not coded in M .

We now show that M fails to be $I\Delta_0$ -full. First notice that for $\theta \in \Pi_1$, if $M \models \theta$ then $\neg \theta \notin T$ so by the maximality of T , $I\Delta_0 + \psi \vdash \theta$ for some $\psi \in T$. (So $M \models \psi$ and $\psi \in \Sigma_1$.) Hence

$$I\Delta_0 \vdash \forall z \exists x \leq z [z \leq 2^z \vee \{(\theta \vee [\theta] \notin x) \wedge (\neg \psi \vee [\theta] \in x)\}]$$

where the conjunction is taken over all pairs $\langle \theta, \psi \rangle$ such that $\theta \in \Pi_1$, $\psi \in \Sigma_1$ and there is a proof with code $\leq n$ of $\psi \rightarrow \theta$ from $I\Delta_0$. To see this consider an arbitrary model of $I\Delta_0$ and if $2^n > z$ take x to be a code for the set of true Π_1 sentences with codes less than n .

However, the corresponding infinitary sentence is not true in M . For otherwise take $z > \mathbb{N}$. Then for some $x \leq z$,

$$\bigwedge_{\langle \theta, \psi \rangle} M \models (\theta \vee [\theta] \notin x) \wedge (\neg \psi \vee [\theta] \in x)$$

which by the above remarks forces x to code $\Pi_1(M)$. It follows that M is not $I\Delta_0$ -full since clearly in the definition of $I\Delta_0$ -full we can equivalently take the θ_i to be Π_1 rather than Δ_0 .

We now show that, assuming $I\Delta_0 \vdash \neg \Delta_0 H$, M has no proper end extension to a model of $I\Delta_0$. For suppose on the contrary $M \subsetneq M' \models I\Delta_0$. Pick $s \in M' - M$ and, using the notation of the proof of Theorem 5(3) let A be the set of (codes of) Σ_1 formulae $\exists \vec{x} \theta(\vec{x})$, $\theta(\vec{x}) \in \Delta_0$ in the sense of M' such that in M' ,

$$\exists \vec{a} <_s [\xi_\theta(\vec{a}) \leq s \wedge \models_s \theta(\vec{a})].$$

Since A is Δ_0 definable in M' , $A \cap \mathbb{N}$ is coded in M' and hence in M since $\mathbb{N} \subset M \subset M'$. We claim that $A \cap \mathbb{N} = \Sigma_1(M)$. To see this, notice that if $M \models \exists \vec{x} \theta(\vec{x})$ with $\theta(\vec{x}) \in \Delta_0$ then $M \models \theta(\vec{a})$ for some $\vec{a} \in M$ and $\xi_\theta(\vec{a}) \in M < s$ and $\models_s \theta(\vec{a})$ since $M' \models \theta(\vec{a})$. Conversely, if $[\exists \vec{x} \theta(\vec{x})] \in A \cap \mathbb{N}$ then $M' \models \exists \vec{x} \theta(\vec{x})$ and hence, by the maximality of T , $M \models \exists \vec{x} \theta(\vec{x})$.

Thus $\Sigma_1(M)$, and hence $\Pi_1(M)$, is coded in M giving the required contradiction. \square

REMARKS. In attempting to answer the question “under what (natural) conditions does a countable model K of $I\Delta_0$ have a proper end extension to a model of $I\Delta_0$ ”, we have suggested two possible answers:

- (i) $K \models B\Sigma_1$;
- (ii) K is $I\Delta_0$ -full and $K \models B\Sigma_1$.

The former is necessary, the latter sufficient, so the “true” answer (assuming there is one) lies somewhere between these extremes. Under the further restriction on K that K is closed under exponentiation (i) is both necessary and sufficient whilst under the assumption $I\Delta_0 \vdash \neg \Delta_0 H$ (i) is not sufficient and for a wide class of K , (ii) is necessary and sufficient.

We remark here that what is required is a “natural” condition so it may well be that the problem has no satisfactory answer. Certainly there is an unsatisfactory answer for it is not difficult to write down an infinitary sentence X such that for countable $K \models I\Delta_0 + B\Sigma_1$, K has a proper end extension to a model of $I\Delta_0$ if and only if $K \models X$. However, the sentence X only tells us what we already know and gives no new insight into the problem.

Finally, we remark here that the sufficiency of $K \models B\Sigma_1$ is related to another intriguing open question of whether every model of $I\Delta_0 + \neg B\Sigma_1$ must be closed under exponentiation, i.e. whether

$$I\Delta_0 + \neg \forall x \exists y y = 2^x \vdash B\Sigma_1 .$$

The relationship is that one of these must fail. To see this, suppose both hold and take $M_0 \models I\Delta_0$ in which there is a semantic tableaux proof of $0 \neq 0$ from $I\Delta_0$. Then by results of PARIS and WILKIE (1987), M_0 cannot be closed under exponentiation so $M_0 \not\models B\Sigma_1$ and hence M_0 has a proper countable end extension M_1 , $M_1 \models I\Delta_0$. The semantic tableaux proof is still in M_1 so again $M_1 \not\models B\Sigma_1$ and hence we can find $M_1 \subsetneq M_2$, $M_2 \models I\Delta_0$. Continuing in this way ω_1 times gives an ω_1 -like model of $I\Delta_0$ which, because it must satisfy $B\Sigma_n$ for all n , must be a model of Peano's Axioms. However, this is a contradiction since it still contains the original semantic tableaux proof of $0 \neq 0$ from $I\Delta_0$!

References

- BOUGHATTAS, S., Thèse 3^{ème} Cycle, Université Paris VII, to appear.
- DIMITRACOPOULOS, C. and PARIS, J., 1982, *Truth definitions for Δ_0 formulae*, Logic and Algorithmic, Monographie No. 30 de L'Enseignement Mathématique, Genève, pp. 319–329.
- DIMITRACOPOULOS, C. and PARIS, J., *A note on a theorem of H. Friedman*, to be submitted.
- FRIEDMAN, H., 1971, *Countable models of set theories*, LNM, Vol. 337 (Springer-Verlag), pp. 539–573.
- KIRBY, L. and PARIS, J., 1978, Σ_n -collection schemes in arithmetic, Logic Colloquium, 1977, North-Holland, pp. 199–209.
- LESSAN, H., 1978, Ph.D. thesis, Manchester University.
- MACDOWELL, R. and SPECKER, E., 1961, *Modelle der Arithmetik*, Infinitistic Methods, Proc. Symp. on the Foundations of Math., Warsaw 1959, Oxford, pp. 257–263.
- PARIS, J., 1981, *Some conservation results for fragments of arithmetic*, LNM, Vol. 890 (Springer-Verlag), pp. 251–262.
- PARIS, J. and WILKIE, A., 1987, *On the scheme of induction for bounded arithmetic formulas*, Ann. Pure Appl. Logic 35, pp. 261–302.
- SHEPHERDSON, J., 1965, *Non-standard models for fragments of number theory*, The Theory of Models (North-Holland), pp. 342–358.

This Page Intentionally Left Blank

TOWARDS THE STRUCTURAL STABILITY THEORY

B.I. ZILBER

Department of Mathematics, Kemerovo State University, Kemerovo, USSR

In recent years, many works in stability theory have appeared which are more or less connected with the trichotomy in the class of uncountably categorical theories discussed by ZILBER (1984a) and extended to the class of stable theories (PILLAY 1984, BUECHLER 1986a). Here I am going to give a survey of these works; the article (ZILBER 1984a) will be used as the basic one.

1. Weak normality

Recall that the question whether or not a pseudoplane is definable in an uncountably categorical structure is of principal significance. Pillay introduced a closely related criterion, the notion of (weak) normality (PILLAY 1984).

A definable subset S of a saturated structure M is said to be (weakly) normal if for any $a \in M$ there is at most one (finite number of) conjugate(s) of S containing a . The theory of M and M are said to be (weakly) normal if any X -definable subset of M^{c^q} is a Boolean combination of $\text{acl}(X)$ -definable (weakly) normal subsets.

Buechler introduced an equivalent notion of 1-basedness (BUECHLER 1986a).

It is easy to prove that weak normality is equivalent to the non-definability of pseudoplanes, provided we are in the class of uncountably categorical theories. For the class of stable theories this is not true; in (BELEGRADEK 1987) it is shown that the free groupoid with two generators has a normal theory and at the same time a pseudoplane is definable in it. Nevertheless, the two approaches are in fact equivalent. Pillay proved in (HRUSHOVSKI and PILLAY 1987b) that non-weakly normal stable structures

are exactly those which have no type-interpretable pseudoplane. Here the type-interpretation of a pseudoplane means that the sets of points and lines and the incidence relation are all defined by complete types over some set.

In (ZILBER 1984a) the class of weakly normal uncountably categorical structures was divided into two subclasses: the class of module-like structures and that of disintegrated type structures. The classes were singled out by types of geometries associated with strongly minimal sets in the structures. Module-like structures have locally modular non-degenerate (locally projective) geometries, structures of disintegrated type have trivial geometries of strongly minimal sets. Since geometries can be associated with strongly minimal and even regular type-definable sets, too, the characteristics can be used in the stable case. HRUSHOVSKI (1985) has shown that if the geometry of a regular type in a stable M is locally modular non-degenerate then in M an infinite Abelian group is definable, thus justifying the term "module-like".

It is important to clear up the connection of the property of having locally modular geometries for all regular types and weak normality. A partial result here is known: the arguments which establish the connection of the definability of a pseudoplane in a strongly minimal structure with the geometry of the structure (ZILBER 1986), used with U-rank instead of Morley rank, fit to prove

THEOREM 1. *The geometries of strongly minimal type-definable sets in weakly normal structures are locally modular.*

Unfortunately, these considerations say nothing about the geometries of regular types in weakly normal structures.

BUECHLER (1985) and PILLAY (1985) have proved

THEOREM 2. *If in a superstable theory of finite U-rank all strongly minimal type-definable sets are locally modular, then the theory is weakly normal.*

However, the examples of Hrushovski and Baudisch cited in PILLAY (1985) show that there are superstable theories with all regular types locally modular which are not weakly normal.

In spite of these counterexamples there is a hope that the converse of Theorem 1 is true for stable \aleph_0 -categorical theories. This is of special interest since it is known that strongly minimal types in \aleph_0 -categorical

stable theories are locally modular (see Theorem 6 below), and that weakly normal \aleph_0 -categorical theories are \aleph_0 -stable (LACHLAN 1973/74).

Theorem 2 is in fact a consequence of the important Coordinatization Theorem originating from CHERLIN *et al.* (1985). The theorem, roughly speaking, states that the global structure is well reflected by the structure of a “small” (rank 1) subset. In stable theories the connection between global and local properties can be looser. Therefore the conjecture of ZILBER (1984a) about field-like structures should be restated for the stable case in the following form:

If in a stable theory there is a regular type with non-locally modular geometry then an infinite field is definable in the theory.

Buechler developed the idea of the Coordinatization Theorem (BUECHLER 1985) to prove

THEOREM 3. *If a strongly minimal type-definable set S is a subset of a definable set of ∞ -rank 1 and is not locally modular then S is a subset of a definable strongly minimal set.*

The discussion above shows that the classification of theories from the combinatorial-geometric point of view is interesting for quite a wide range of stable theories. The efficiency of the combinatorial-geometric approach can be well explained by the thesis: a key consequence of stability theory is the ability to define a notion of dimension for a wide class of structures (see BALDWIN 1984). If so then combinatorial geometries (matroids) must play an important role in stability theory. But the deep combinatorial theory begins when separate types of matroids are studied. As a matter of fact, this was confirmed by the Trichotomy Theorem (ZILBER 1984a, 1986) which singled out the three classes of structures: of disintegrated type, module-like and field-like. Below we follow this scheme.

2. Results by Lachlan, Cherlin and Shelah

The starting point of works by LACHLAN (1984a), Lachlan and SHELAH (1984b) and CHERLIN and LACHLAN (1986) is Lachlan's idea that the problem of classification of homogeneous graphs considered by Gardiner fits a suitable model-theoretic scheme.

Fix a finite relational language L . Recall that an L -structure M is said to be (finitely) homogeneous if its theory admits elimination of quan-

tifiers or equivalently for countable M : any isomorphism between two finite substructures of M is induced by some automorphism of M .

It is reasonable to enlarge the family of all finite homogeneous L -structures by adding to it all infinite stable homogeneous L -structures. The resulting class, denoted $\text{Hom}(L)$ is here the object of study.

It is easy to see that any infinite structure M from $\text{Hom}(L)$ has a \aleph_0 -stable, \aleph_0 -categorical theory. More than that, the geometries of strongly minimal sets in M are degenerate, i.e. M is of disintegrated type. The global structure of M can be reduced to the local one using the Coordinatization Theorem. Here a more precise version of the theorem is used. The simplest form of the main result is the Dichotomy Theorem of CHERLIN and LACHLAN (1986).

THEOREM 4. *There is an integer m (depending on L) such that for every $M \in \text{Hom}(L)$ and every maximal O -definable equivalence relation E on M one of the following holds:*

- (A) $|M/E| \leq m$,
- (B) M/E is quasi-coordinatizable by means of Grassmannians.

“Quasi-coordinatizable by means of Grassmannians” can be easily explained when M/E is infinite. In this case there are definable strongly minimal sets S_1, \dots, S_n which are also mutually indiscernible sets and a number k such that we can identify every point of M/E with a set $X \subseteq S_1 \cup \dots \cup S_n$ such that $|X \cap S_i| = k$ for every i . X is called the set of coordinates of the point. The identification is definable; thus it follows that all relations between points from M/E are reduced to coincidences among coordinates.

Of course, the theorem does not give a complete description of M yet, as we do not know what is beyond E . The further analysis allows any M to “shrink” canonically to some finite $M_0 \in \text{Hom}(L)$, where the number of elements of M_0 is bounded by a fixed number depending only on L , the isomorphism type of M is uniquely determined by M_0 and some “dimensions of M ”, cardinalities of some coordinate sets S_i .

The scheme can be finitized, i.e. a finite M can be considered in the same way. But then certain difficulties arise: one must define correctly analogues of strongly minimal sets, indiscernibles and ranks in finite structures. This needs some recent classification results in finite group theory.

In the finite analysis the three papers give an ideal structural theory of

the class $\text{Hom}(L)$ which in some cases, for example for graphs, gives an explicit classification. The results demonstrate the strength of the notions and methods of modern stability theory and their outward applicability.

3. Module-like structures

This case is far more complicated and so results here are not so comprehensive as yet. However, results concerning this class are quite deep and spectacular both from the model-theoretic and the general point of view.

The main objects of study here are the strongly minimal structures and the pregeometries associated with them.

Recall that in ZILBER (1981, 1984b, 1984c) the complete classification of the geometries of strongly minimal \aleph_0 -categorical structures was obtained. The same result was obtained as a consequence of the known classification of finite 2-transitive groups in CHERLIN *et al.* (1985) and by some other mathematicians. D. Evans, a representative of the combinatorial school, using some constructions from ZILBER (1981), gave another, purely combinatorial proof of the theorem (EVANS 1986) and got a finite version of the theorem for a dimension of at least 23 in a still unpublished appendix to EVANS (1986) (see Theorem 5 below). Here the methods of ZILBER (1981, 1984b, 1984c) which have been developed further, will be discussed.

Recall that the proof in ZILBER (1984b, 1984c) is divided into four main steps:

Step 1. The local modularity of a strongly minimal structure V is equivalent to the statement that on any set P of Morley rank 2 definable in V , there does not exist a definable family L of lines (strongly minimal subsets of P) of rank 2, any two of which would have at most finite intersection (non-existence of a pseudoplane).

Step 2. Suppose the pseudoplane (P, L) exists. It is possible (using \aleph_0 -categoricity) to associate a polynomial $p_S(x)$ over rationals to any definable subset $S \subseteq P$. The mapping $S \rightarrow p_S$ has some good computational properties. After carrying out a series of computations one can get the following property of the geometry of V :

for any strongly minimal $S \subseteq V \times V$, there are $a_1, a_2 \in V$ such that S is $\text{acl}(a_1, a_2)$ -definable.

This is called semi-projectivity (or 2-modularity). More generally, if the

property holds for a_1, \dots, a_k for a fixed k instead of a_1, a_2 , it is called k -modularity or pseudo-modularity (Buechler).

Step 3. Using 2-modularity of V one succeeds in constructing another strongly minimal set U , non-orthogonal to V , on which there are "enough" quasi-translations, i.e. definable convertible mappings from U to U .

Step 4. Using the results of steps 2 and 3 one constructs a definable group G of translations that acts generically on U . The Morley rank of G is 2 iff the pseudoplane exists. From facts known for such groups [(BAUR *et al.* 1979) for the \aleph_0 -categorical case and (CHERLIN 1979) for the general one] it is deduced that either the center of G is of rank 1, or an infinite field is definable in G . The first case is possible only if the rank of G is 1, as G acts faithfully on U of rank 1; the second is inconsistent with \aleph_0 -categoricity (and even with pseudo-modularity).

This contradiction completes the proof.

Let us analyse possibilities to generalize the proof.

To reduce the analysis of the geometry on V to the analysis of a pseudoplane-like structure (Step 1) one needs a good rank notion. In an arbitrary homogeneous pregeometry V for $\langle v_1, \dots, v_n \rangle \in V^n$ the rank could be defined as

$\text{rank}(\langle v_1, \dots, v_n \rangle / X) = \dim \text{cl}(v_1, \dots, v_n, X) - \dim \text{cl}(X) =$ the maximal number of mutually X -independent elements of $\{v_1, \dots, v_n\}$.

This definition follows the definition of the rank (dimension of a variety) in algebraic geometry. In general one also needs to consider "imaginary elements" $u \in V^n/E$, where E is an equivalence relation which is X -definable (or $\text{Aut}(V/X)$ -invariant). In fact, it is enough to deal only with relations E which have classes of rank 0 only. For such an E it is natural to put for $u \in \bar{v}/E$, $\bar{v} \in V^n$: $\text{rank}(u/X) = \text{rank}(\bar{v}/X)$.

This definition works well rather generally to achieve the end of Step 1, provided $\dim V \geq 5$.

The same rank notion can be used to carry out Cherlin's analysis of a connected group of rank 2 and to get either a field of rank 1 or a large center in the group (Step 4).

The realization of Step 2 depended on some calculations with polynomials. Observe that the same polynomials can be used when V is a finite pregeometry and $\dim V$ is not less than the degrees of the polynomials involved. Nevertheless, the full use of the polynomials is redundant, as could be easily seen by inspecting (ZILBER 1984b); in fact only the leading

coefficients are considered in the computations. The leading coefficients of polynomials coincide if the corresponding sets almost coincide; thus we can speak of the leading coefficient of a type: $\text{lcp}(u/X)$. It can be computed for an element $u \in V^{\text{eq}}$ such that $\text{cl}(u, X) = \text{cl}(\bar{v}, X)$, $\bar{v} \in V^n$ as

$$\text{lcp}(u/X) = \frac{\text{Mult}(u/\bar{v}X)}{\text{Mult}(\bar{v}/uX)}$$

The definition is correct, as observed by Hrushovski, iff V is unimodular, i.e.

$$\begin{aligned} \text{Mult}(\bar{v}/\bar{v}') &= \text{Mult}(\bar{v}'/\bar{v}) \quad \text{for } \bar{v}, \bar{v}' \in V^n, \quad \text{cl}(\bar{v}) = \text{cl}(\bar{v}'), \\ \dim(\bar{v}) &= \dim(\bar{v}') = n. \end{aligned}$$

Hrushovski carried out a very interesting analysis of the property and of the function lcp in HRUSHOVSKI (1987a). In particular, he made the following deep observation: if we put $\text{deg } S = \text{lcp } S/\sqrt{\text{lcp } P}$ for subsets S of P (the points of the pseudoplane), then the computations from ZILBER (1984b) lead, in fact, to the ‘‘Bezout formula’’ for intersections of two curves (subsets of rank 1):

$$|S_1 \cap S_2| = \text{deg } S_1 \cdot \text{deg } S_2.$$

The formula is proved under the condition S_1 is an element of V^{eq} of rank 1, S_2 an element of rank 2 and the elements are independent. Hrushovski shows that if the formula were right for all independent S_1, S_2 , then one would get the final contradiction without using Steps 3 and 4. But under the present conditions one gets only 2-modularity.

Observe that $\text{acl} = \text{dcl}$ on strongly minimal set V implies that V is unimodular and $\text{lcp}(S) = 1$ for all strongly minimal subsets $S \subseteq P = V \times V$ and $\text{lcp}(P) = 1$, too. Thus the counting arguments of Step 2 are trivial in this case and may be eliminated, as was done in ZILBER (1987e) in the proof of Theorem 7 presented below.

Step 3 was the longest and most unnatural one in ZILBER (1984c). Recently the author and Hrushovski found independently different ways to construct quasi-translations and thus to get the group for Step 4. Both constructions allow broad generalizations and this led to new results. Hrushovski’s proof goes for regular types and is presented in the draft paper by HRUSHOVSKI (1987a). I take the opportunity to present here the crucial point of my construction.

In the Lemma below let V be a strongly minimal geometry, i.e. $\text{acl}(v) = \{v\}$ for all $v \in V$.

LEMMA. Let $f, h, g \subseteq V \times V$ be strongly minimal and $h \subseteq f \cdot g$ (the composition of binary relations). As elements of V^{eq} let

$$\text{rank}(f/\emptyset) = \text{rank}(h/\emptyset) = \text{rank}(g/\emptyset) = k ;$$

$$\text{tp}(f/\emptyset) = \text{tp}(h/\emptyset); f, g \text{ are independent .}$$

Under these conditions there is for every $v \in V - \text{acl}(f, g)$ and $u \in f(v)$ a unique $w \in g(u) \cap f(v)$. (Here $u \in f(v)$ means $\langle v, u \rangle \in f$ and so for g, h, \dots)

PROOF. Let us denote

$$|f(v)| = m, \quad f(v) = \{u_1, \dots, u_m\}, \quad |g(u_i) \cap f(v)| = m_i.$$

Since the subset of f

$$f' = \{\langle v', u' \rangle \in f : \exists w' \in g(u') \cap h(v')\}$$

is of rank 1, it almost coincides with f and so $\langle v, u_i \rangle \in f$ for every $u_i \in f(v)$. Hence $m_i \geq 1$. From the independence of v, f, g it follows easily that $g(u_i) \cap g(u_j) = \emptyset$ for $i \neq j$. Hence $m = |h(v)| = \sum_{i \leq m} m_i$, thus $m_i = 1$ for all $i \leq m$. \square

This lemma gives a definable mapping t_{fh} from the set f to the set h defined as

$$t_{fh} : \langle v, u \rangle \rightarrow \langle v, w \rangle, \quad w \text{ being the only element of } g(u) \cap f(v).$$

In the same way we can consider $t_{h^{-1}f^{-1}}$ from h^{-1} to f^{-1} . Let i denote the inversion on $V \times V$,

$$i : \langle v_1, v_2 \rangle \rightarrow \langle v_2, v_1 \rangle, \quad \text{and put } t = i \cdot t_{h^{-1}f^{-1}} \cdot i \cdot t_{fh}.$$

It is easy to see that t defines a bijection, except for a finite number of points, from f onto itself, and for different h we get different t . These t are the quasi-translations on the set $U = f$.

REMARK. In our case $k = 2$ (V is 2-modular by Step 2). One chooses independent f, h , at first, and then $g \subseteq f^{-1} \cdot h$.

Note that in the more general case, if $\text{acl}(v) \neq \{v\}$, one should find a definable equivalence relation E such that on V/E , $g(u_i) \cap g(u_j) = \emptyset$ could be proved (this is the only use of the condition). In the case in which V is a strongly minimal X -type-definable set in an \aleph_0 -categorical theory, E could be defined as

$$v_1 E v_2 \Leftrightarrow \text{acl}(v_1, X) = \text{acl}(v_2, X) \ \& \ \text{acl}(v_1, h) = \text{acl}(v_2, h)$$

for some h realizing $\text{tp}(f/X)$; h and v_1 independent over X . E depends only on $\text{tp}(f/X)$. On the other hand E is definable, since $\text{acl}(v, h)$ is finite.

Note also that when $\text{acl} = \text{dcl}$ on V , Step 3 is redundant since in this case $u \in \text{acl}(v, X) - \text{acl}(X)$ implies the existence of an X -definable bijection $t : V - \text{acl}(X) \rightarrow V - \text{acl}(X)$ such that $f(v) = u$.

The analysis above allows broad applications. In this way the following results were obtained.

(i) *Finite geometries.* The theorem for $\dim V \geq 23$ was proved by Evans (see the beginning of the section). The arguments presented above give the present form (Zilber).

THEOREM 5. *Let V be a finite homogeneous geometry with at least three points on its lines, $\dim V \geq 7$. Then V is a projective or affine geometry over a finite field.*

(ii) *Strongly minimal sets in stable \aleph_0 -categorical theories.* HRUSHOVSKI (1987a) and ZILBER (1987) independently.

THEOREM 6. *Let V be an X -type-definable strongly minimal set in a model of a stable \aleph_0 -categorical theory. Then the geometry of V over X is locally modular.*

(iii) *Hereditarily transitive groups.* This is a generalization of the case $\text{acl} = \text{dcl}$ discussed in ZILBER (1984a).

Let G be a permutation group on the set V . For any $X \subseteq V$ put

$$G_X = \{ g \in G : \forall x \in X \ gx = x \},$$

$$\text{cl}(X) = \{y \in V : \forall g \in G_X \text{ } gy = y\}.$$

We call G hereditarily transitive if for every $X \subseteq V$ the group G_X acts transitively on $V - \text{cl}(X)$.

It is easy to check that (V, cl) is a pregeometry.

EXAMPLES: 1. If V is an X -type-definable strongly minimal set and $\text{acl}(X \cup Y) \cap V = \text{dcl}(X \cup Y) \cap V$ for every $Y \subseteq V$ then $\text{Aut}(V/X)$ is a hereditarily transitive group.

2. Automorphism groups of v^* -algebras (see NARKIEWICZ (1962)) are hereditarily transitive.

The proof of the following is in ZILBER (1984d).

THEOREM 7. *If G is a hereditarily transitive group on V , $\dim V = n \geq 8$ and there are $v_1, \dots, v_{n-4} \in V$ such that $\text{cl}(v_1, \dots, v_{n-4}) \neq \text{cl}(v_1) \cup \dots \cup \text{cl}(v_{n-4})$, then G is a "large" subgroup of*

$$GL_W(\bar{V}) = \{g \in GL(\bar{V}) : g \text{ is the identity on } W\}$$

or

$$AGL_W(\bar{V}) = T(\bar{V}) \cdot GL_W(\bar{V}),$$

where \bar{V} is a vector space over a division ring; the sets \bar{V} and V differ only in points fixed by the corresponding groups; $W \subseteq \bar{V}$, $T(\bar{V})$ is the group of all translations of V . "Large" subgroup means that for any two n -tuples of linearly (affinely) independent points of \bar{V} there is a unique $g \in G$ bringing the first onto the second.

If V is finite then $G = GL_W(\bar{V})$ or $G = AGL_W(\bar{V})$. The same is true if V is a v^* -algebra and $G = \text{Aut}(V)$.

(iv) *Unimodularity.* The definition (see above) and the result is by HRUSHOVSKI (1987a).

THEOREM 8. *If V is a strongly minimal unimodular structure then the geometry of V is locally modular.*

4. What to do in the field-like case?

Unfortunately, there is practically no progress in this case. After works on module-like and disintegrated type structures one can rather safely assume that any progress in studying this class would be connected with new knowledge in the associated field of mathematics, this time in algebraic geometry (over algebraically closed fields).

First of all natural field-like structures should be studied, and so I would like to attract once again attention to the following problem:

Show that in any structure M definable in an algebraically closed field F the field is definable, provided M is rich enough to interpret a pseudo-plane.

The solution of the problem is known in very special cases, but even the following easy result: “in any simple algebraic group G over an algebraically closed field F the field is definable” has a rather serious corollary: any group automorphism of G is the composition of a rational automorphism and of an automorphism induced by an automorphism of the field (POIZAT 1985, ZILBER 1984d). This is the Borel–Tits theorem restricted to algebraically closed fields (BOREL and TITS 1973).

If the main conjecture on field-like structure (ZILBER 1984a) is true then any strongly minimal field-like structure is equivalent, in a known sense, to a field. But since we are far from proving it, it seems reasonable to look for axiom-like conditions on a homogeneous pregeometry which imply that the pregeometry is induced by an algebraically closed field. The following two conditions could possibly be used in an axiomatization of these pregeometries.

(1) *Parametrization of degree.* There exists a function f of two integer variables to positive integers such that if $y \in \text{cl}(X, z) - \text{cl}(X)$ then there is $X' \subseteq \text{cl}(X)$ with $|X'| \leq f(\text{Mult}(y/Xz), \text{Mult}(z/Xy))$ and $y \in \text{cl}(X', z)$.

If in the condition f is constant then this is exactly pseudo-modularity. BUECHLER (1986b) and HRUSHOVSKI (1987a) proved that pseudo-modularity implies local modularity of a pregeometry. Thus one must assume f cannot be chosen bounded. The words “parametrization of degree” are to underline the connection with the following fact: any algebraic curve of a given degree on the plane can be parametrized by a fixed number of parameters from the field, depending only on the degree. On the other hand there is an obvious connection between the pair $(\text{Mult}(y/Xz), \text{Mult}(z/Xy))$ and the degree of the curve defined by (y, z) over X , at least in the characteristic 0 case. When the characteristic is positive the connection could be complicated a little by Frobenius automorphisms.

(2) *Quasi-finiteness*. If g is a definable injection of a definable set S into itself then g is a bijection.

This property of algebraically closed fields is in fact proved by CHERLIN (1976).

References

- BALDWIN, J.T., 1984, *Strong saturation and the foundation of stability theory*. Logic Colloq. 82. Proc. Colloq., Florence, 23–28 Aug., 1982 (North-Holland, Amsterdam), pp. 71–84.
- BAUR, W., CHERLIN, G. and MACINTYRE, A., 1979, *Totally categorical groups and rings*, J. Algebra, 57, pp. 407–440.
- BELEGRADEK, O.V., 1988, *Model theory of locally free algebras*, in: Y.L. Ershov, ed., Model Theory. Proc. Inst. Math. Siberian Branch of Ac.Sc.USSR. Novosibirsk, pp. 3–25.
- BOREL, A. and TITS, J., 1973, *Homomorphisms “abstrats” de groupes algébriques simples*, Ann. Math., 97, pp. 499–571.
- BUECHLER, S., 1985, *Coordinatization in superstable theories I. Stationary types*, Trans. AMS, 288 (1), pp. 101–114.
- BUECHLER, S., 1985a, *“Geometrical” stability theory*, Preprint.
- BUECHLER, S., 1986b, *Pseudo-modular strongly minimal structures*, Preprint.
- BUECHLER, S., *Locally modular theories of finite rank*, Ann. Pure and Appl. Logic (to appear).
- CHERLIN, G., 1976, *Model theoretic algebra. Selected topics*, Lect. Notes in Math. 521.
- CHERLIN, G., 1979, *Groups of small Morley rank*, Ann. Math. Logic, 17, pp. 1–28.
- CHERLIN, G., HARRINGTON, L. and LACHLAN, A.H., 1985, \aleph_0 -categorical, \aleph_0 -stable structures, Ann. Pure and Appl. Logic, 28 (2), pp. 103–136.
- CHERLIN, G. and LACHLAN, A.H., 1986, *Stable finitely homogeneous structures*, Trans. AMS, 296, pp. 815–850.
- EVANS, D., 1986, *Homogeneous geometries*, Proc. London Math. Soc., 52, pp. 305–327.
- HRUSHOVSKI, E., 1985, *Locally modular regular types*, Preprint.
- HRUSHOVSKI, E., 1987a, *Recognizing group action in Jordan geometries*, Preprint.
- HRUSHOVSKI, E. and PILLAY, A., 1987b, *Weakly normal groups*, in: The Paris Logic Group, eds., Logic Colloq '85 (Elsevier Sci. Publ.), pp. 121–130.
- LACHLAN, A.H., 1973/74, *Two conjectures regarding the stability of ω -categorical theories*, Fund. Math., 81, pp. 133–145.
- LACHLAN, A.H., 1984a, *On countable stable structures which are homogeneous for a finite relational language*, Isr. J. Math., 49, pp. 69–153.
- LACHLAN, A.H. and SHELAH, S., 1984b, *Stable structures homogeneous for a finite binary language*, Isr. J. Math., 49, pp. 155–180.
- NARKIEWICZ, W., 1962, *Independence in a certain class of abstract algebras*, Fund. Math., 50, pp. 333–340.
- PILLAY, A., 1984, *Stable theories, pseudoplanes and the number of countable models*, Preprint.
- PILLAY, A., 1985, *Simple superstable theories*, Preprint.
- POIZAT, B., 1988, *Mrs. Borel, Tits, Zilber et le nonsense général*, J Symb. Logic, 53, pp. 124–131.
- ZILBER, B.I., 1981, *Totally categorical structures and combinatorial geometries*, Doklady Ac.Sc.USSR, 259, pp. 1039–1041 (in Russian).

- ZILBER, B.I., 1984a, *The structure of models of uncountably categorical theories*, in: P. Olech, ed., Proc. ICM-83 (PWN—North-Holland, Warszawa, Amsterdam) pp. 359–368.
- ZILBER, B.I., 1984b, *Strongly minimal countably categorical theories II*, Siberian Math. J., 25(3), pp. 396–412 (in Russian).
- ZILBER, B.I., 1984c, *Strongly minimal countably categorical theories III*, Siberian Math. J., 25(4), pp. 559–571 (in Russian).
- ZILBER, B.I., 1984d, *Some model theory of simple algebraic groups over algebraically closed fields*, Colloq. Math. 48(2), pp. 173–180.
- ZILBER, B.I., 1986, *Structural properties of models of \aleph_1 -categorical theories*, in: P. Weingartner, ed., Proc. 7-th Int. Congr. LMPS. Salzburg, 1983 (North-Holland, Amsterdam), pp. 115–128.
- ZILBER, B.I., 1987, *On the dichotomy for stable theories: weak normality and pseudoplanes*, in: I.T. Frolov, ed., Abstracts of Int. Congr. LMPS, Moscow, 1987, vol. 5, Part 1 (Nauka, Moscow) pp. 92–93.
- ZILBER, B.I., 1988, *Hereditarily transitive groups and quasi-Urbanik structures*, in: Y.L. Ershov, ed., Model Theory. Proc. Inst. Math. Siberian Branch of USSR Ac.Sc. Novosibirsk, pp. 58–77.

This Page Intentionally Left Blank

3

**Foundations of Computing
and Recursion Theory**

This Page Intentionally Left Blank

AUTOMORPHISMS OF THE LATTICE OF RECURSIVELY ENUMERABLE SETS AND HYPERHYPERSIMPLE SETS

EBERHARD HERRMANN

Sektion Mathematik, Humboldt-University, DDR-1086 Berlin, GDR

After the proof of Soare that all maximal sets are automorphic in (SOARE 1974) the next greater class of r.e. sets analyzed from this point of view was and is the class of hyperhypersimple sets.

Concerning the automorphism properties of these sets, some facts are already known. The results in question were shown by HERRMANN (1983, 1986), LERMAN *et al.* (1984) and MAASS (1984).

While MAASS (1984) proved a sufficient criterion for hh-simple sets to be automorphic, thus extending the result of Soare, in HERRMANN (1983) those properties of these sets were analyzed, from which it can be concluded when such sets are not automorphic (even if there r.e. superset structures are isomorphic). This extends the result of LERMAN *et al.* (1984).

For the presentation of the lattice properties of the hh-simple sets the following tool is very useful:

Let $2^{<\omega}$ be the set of all finite sequences of 0's and 1's, which we shall call "words". $\langle \rangle$ denotes the empty word and $|r|$ the length of the word r . If a and b are from $2^{<\omega}$ we denote with $a * b$ the concatenation of both (in particular, we have $a * 0$ and $a * 1$), $a \leq b$ means that a is an initial part of b ($a < b$ that $a \leq b \wedge a \neq b$). With $<_l$ we denote the lexicographical order in $2^{<\omega}$ (see LACHLAN 1968, p. 16).

Special subsets of $2^{<\omega}$ are for us of particular interest: A subset Γ of $2^{<\omega}$ is called *branch* if $\Gamma \neq \emptyset$ and if $a \in \Gamma$ then $b \leq a \rightarrow b \in \Gamma$ and either $a * 0 \in \Gamma$ or $a * 1 \in \Gamma$. With $[a]$, for $a \in 2^{<\omega}$ we denote the set $\{b \in 2^{<\omega} : a \leq b\}$.

A subset Δ of $2^{<\omega}$ is called *ideal* if

$$a \in \Delta, b \in 2^{<\omega} \rightarrow a * b \in \Delta$$

and

$$a * 0 \in \Delta, a * 1 \in \Delta \rightarrow a \in \Delta.$$

An ideal Δ is called Σ_3^0 -ideal if Δ is a Σ_3^0 -set (by any effective coding of the elements of $2^{<\omega}$ by numbers). Just these ideals are of particular interest for us. Let E_1^3 be the family of all Σ_3^0 -ideals of $2^{<\omega}$ together with the operations \vee ($\Delta_1 \vee \Delta_2$ is the smallest ideal including both. If Δ_1, Δ_2 are from Σ_3^0 then also $\Delta_1 \vee \Delta_2$) and \wedge ($\Delta_1 \wedge \Delta_2$ is the intersection of both sets). Let E_1^{3-} be the set E_1^3 without the ideal $2^{<\omega}$. Denote by $E_{1,r}^3$ the sublattice formed by all elements of E_1^3 with a complement in E_1^3 (respectively to the ideals \emptyset and $2^{<\omega}$). Let $\Delta_0 \in E_1^3$. By $E_1^3(\Delta_0)$ ($E_{1,r}^3(\Delta_0)$) we denote the sublattice

$$\{\Delta \in E_1^3 : \Delta_0 \leq \Delta\} \quad (\{\Delta_0 \vee \Delta' : \Delta' \in E_{1,r}^3\}).$$

The usefulness of the notions of Σ_3^0 -ideals and the sublattices $E_1^3(\Delta)$ and $E_{1,r}^3(\Delta)$ for the analysis of the hh-simple sets shows the following:

“A Boolean algebra \mathfrak{A} is an $\exists\forall\exists$ -Boolean algebra iff there is a $\Delta \in E_1^{3-}$ such that \mathfrak{A} and $E_{1,r}^3(\Delta)$ are isomorphic” (see HERRMANN 1983).

Hence the isomorphism types of $L^*(A)$, A hh-simple and $E_{1,r}^3(\Delta)$, $\Delta \in E_1^{3-}$ coincide.

Later this can still be improved (see Theorem 1).

Of fundamental importance for our paper is the general construction of hh-simple sets carried out by LACHLAN (1968, p. 21). Thus we use the sequence $(\Psi_{r,s})_{r \in 2^{<\omega}, s \geq 0}$, the functions T_s from $2^{<\omega}$ onto $\overline{A_s}$ and T the (partial) limit function. Additionally to A we define r.e. sets A_r , $r \in 2^{<\omega}$. Let A_r be the set

$$\{x : (\exists s)(T_s(r) = x \wedge x \in A_{s+1})\},$$

i.e. x comes into A from the place r .

The sets A_r have the following important properties:

- 1° $(A_r)_{r \in 2^{<\omega}}$ is an r.e. sequence of disjoint subsets of A with $\bigcup A_r = A$.
- 2° For every $r \in 2^{<\omega}$ the set $A_r^0 \cup T[r]$ is a recursive set ($A_r^0 =_{df} \bigcup \{A_a : r \leq a\}$).
- 3° $T[r]$ is finite iff for almost all a with $r \leq a$ $\lim_s \Psi_{a,s} = \omega$
iff A_r^0 is recursive (this follows from 2°).

From MAASS (1984) it can be shown that every hh-simple set is automorphic to an hh-simple set constructed in Theorem 6 in LACHLAN (1968).

Let B be an hh-simple set and Δ the set

$$\{(e_0, \dots, e_{n-1}) \in 2^{<\omega} : \bar{B} \cap W_0^{e_0} \cap \dots \cap W_{n-1}^{e_{n-1}} = \emptyset, \\ \text{where } W_i^0 = W_i, W_i^1 = \bar{W}_i\}.$$

Then Δ such defined is an Σ_3^0 -ideal and $L^*(B) \cong E_{1,r}^3(\Delta)$. Let A be an hh-simple set constructed as in Theorem 6 in LACHLAN (1968) by using $(\Psi_{r,s})_{r \in 2^{<\omega}, s \geq 0}$ with $\Psi_{r,s}$ such that

$$\{r : (\exists n)(\forall a)(r \leq a \wedge |a| = n. \rightarrow \lim_s \Psi_{r,s} = \omega)\} \tag{1}$$

is just Δ .

The mapping

$$(B \cup W_e)^* \rightarrow \left(A \cup \cup \{T[e * 0] : |e| = e\} \right)^*$$

is an isomorphism between $L^*(B)$ and $L^*(A)$ (both are isomorphic to $E_{1,r}^3(\Delta)$) and is presented by an Σ_3^0 -permutation. The relation

$$A \cup W_f =^* A \cup \cup \{T[e * 0] : |e| = e\}$$

is Σ_3^0 in e and f . Hence, by MAASS (1984), the sets A and B are automorphic.

Thus Lachlan's construction of hh-simple sets is universal from the point of view of their lattice position. Further we see that for the lattice properties of an hh-simple set from the sequence $(\Psi_{r,s})$ only the set (1) is important. We denote with A_Δ an hh-simple set constructed as in LACHLAN (1968) by means of some $(\Psi_{r,s})$ with Δ equal to the set (1).

REMARK. In LACHLAN (1968) it is announced that the construction in Theorem 6 can be done such that the hh-simple set belongs to an arbitrary given high r.e. T-degree. Since all hh-simple sets are dense simple (Martin), the universality of the hh-simple set construction gives

COROLLARY 1. *The T-degrees of an orbit of an hh-simple set are just all high r.e. T-degrees.*

Ideal families of recursively enumerable supersets

In general the isomorphism type of $L^*(A)$ is not a sufficient property for determining the orbit of A as first shown by LERMAN *et al.* (1984). Thus, to find such a property we have to investigate further lattice properties connected with an hh-simple set, which will be done in the following.

Let A be an hh-simple set. To every r.e. subset X of A it is possible to assign a family of r.e. supersets of A , denoted by $I_A(X)$, by

$$\{Y \in L(A) : X \cup (Y \setminus A) \text{ is r.e.}\}.$$

$I_A(X)$ obviously forms an ideal in $L(A)$.

Let $P_1(A)$ be the family of all these ideals, when X varies over all r.e. subsets of A .

What can be said about $P_1(A)$ for an hh-simple set A ? Let $I_A^*(X) = \{Y^* : Y \in I_A(X)\}$, $P_1^*(A) = \{I_A^*(X) : I_A(X) \in P_1(A)\}$. Easy to see is that the mapping

$$(A \cup R)^* \in L^*(A) \rightarrow I_A^*(A \cap R) \in P_1^*(A), \text{ where } R \text{ is a recursive set,}$$

is an isomorphism between $L^*(A)$ and all principal ideals inside $P_1^*(A)$. Are still other ideals in $P_1^*(A)$? Is $P_1(A)$ a lattice under inclusion?

These and other questions can be answered by using the Theorem which now will be proved.

If \mathfrak{A} and \mathfrak{B} are structures of the same language, we write $\mathfrak{A} \cong \mathfrak{B}$ if they are isomorphic.

Let \mathfrak{X}_0 and \mathfrak{X}_1 be classes of structures, all of the same language. We say that \mathfrak{X}_0 and \mathfrak{X}_1 are *isomorphic-equal* if both classes include the same isomorphism types.

THEOREM 1 (HERRMANN 1983). *The classes $\{(P_1^*(A), \leq) : A \text{ hh-simple}\}$ and $\{E_1^3(\Delta) : \Delta \in E_1^{3^-}\}$ are isomorphic-equal.*

PROOF. By the universality of the Lachlan construction it is sufficient to consider only hh-simple sets A_Δ for $\Delta \in E_1^{3^-}$. We show that $(P_1^*(A_\Delta), \leq) \cong E_1^3(\Delta)$. (In the following we write only A for A_Δ .)

Let X be an r.e. subset of A . We assign to X the set Δ_X defined by

$$\{r \in 2^{<\omega} : X \cup T[r] \text{ is r.e.}\}.$$

(i) $\Delta_x \in E_1^3(\Delta)$. Δ_x is an ideal. Suppose $r*0 \in \Delta_x$ and $r*1 \in \Delta_x$. Since $T[r] = {}^*T[r*0] \cup T[r*1]$, also $X \cup T[r]$ is r.e. Hence $r \in \Delta_x$.

Suppose $r \in \Delta_x$. Then $X \cup T[r]$ is r.e. Thus also $(X \cup T[r]) \cap (A_r^0 \cup T[r])$. But this is equal to $(X \cap A_r^0) \cup T[r]$. From this we see that $((X \cap A_r^0) \cup T[r]) \cap (A_{r*0}^0 \cup T[r*0])$ is r.e. and equal to $(X \cap A_{r*0}^0) \cup T[r*0]$. Hence $X \cup T[r*0]$ is r.e. This gives $r*0 \in \Delta_x$. Similarly we show that $r*1 \in \Delta_x$.

$\Delta \subseteq \Delta_x$. Since for $r \in \Delta$ $T[r]$ is finite, in this case obviously $X \cup T[r]$ is r.e. Thus $r \in \Delta_x$. $\Delta_x \in \Sigma_3^0$. We have $r \in \Delta_x$ iff $X \cup T[r]$ is r.e. iff

$$(\exists e)(W_e\text{-recursive, } W_e \subseteq A, A_r^0 \subseteq X \cup W_e). \tag{2}$$

The second equivalence holds by 2°. But this is an Σ_3^0 -definition. Thus Δ_x is an Σ_3^0 -set.

(ii) The mapping $I^*(X) \rightarrow \Delta_x$ is an embedding. Obviously for r.e. subsets X and Y of A with $X \subseteq Y$ we have $\Delta_x \subseteq \Delta_y$. Further, since $L^*(A)$ and $E_{1,r}^3(\Delta)$ are isomorphic, we get $I^*(X) = I^*(Y) \leftrightarrow \Delta_x = \Delta_y$.

(iii) The mapping in (ii) is surjective. This is the difficult part of the proof of the Theorem. For showing this we need two Lemmata. The first one is similar to Theorem 6 (ROGERS 1972, p. 421) and the second one is a generalization of Theorem 12 (ROGERS 1972, p. 313).

Now we formulate these Lemmata and prove them later. In Lemma 1 we use the following convention: For $r \in 2^{<\omega}$ and a number n with $|r| \leq n$ $[r, n]$ denotes the set

$$\{a \in 2^{<\omega} : r \leq a \wedge |a| = n\}.$$

LEMMA 1. Let Δ be from E_1^3 . Then there is an r.e. subset W of $2^{<\omega}$ such that for all r

$$r \in \Delta \rightarrow [r] \subset {}^*W, \tag{3}$$

$$r \notin \Delta \rightarrow (\exists n)([r, n] \cap W = \emptyset). \tag{4}$$

COROLLARY 2. Let Δ and W as in Lemma 1. If $r \notin \Delta$ and Γ is a branch with $r \in \Gamma$ and $\Delta \cap \Gamma = \emptyset$ then $\Gamma \cap W$ is infinite.

PROOF. If $\Gamma \cap W$ would be finite then there exists an m such that for a with $r \leq a$, $m < |a|$, $a \in \Gamma$ it follows $a \in W$. But for $a \in \Gamma$ we have $a \notin \Delta$. Hence by (4) there is an n such that $[a, n] \cap W = \emptyset$. Both together isn't possible. Thus $\Gamma \cap W$ must be infinite.

- Before we formulate Lemma 2 we show the following: Let $Y \subset_{sm} A$.
- Then $A_r^0 \subset^* Y$ iff $r \in \Delta$.
 - ← Is obvious, since $r \in \Delta$ implies that A_r^0 is recursive. Thus from $Y \subset_m A$ we get $A_r^0 \subset^* Y$.
 - If $A_r^0 \subset^* Y$, then $(A_r^0 \cup T[r]) \cap (A \setminus Y) = \emptyset$. Thus $T[r] \cup \emptyset$ is r.e., that means that $T[r]$ is finite. Hence $r \in \Delta$.
 - Let X be an r.e. subset of A and $r \in 2^{<\omega}$. $X \cup T[r]$ is r.e. iff $A_r^0 \subset^* X \cup Y$.
 - Is obvious, since $A_r^0 \cup T[r]$ is r.e.
 - ← $X \cup T[r]$ implies $A_r^0 \subseteq X \cup R$, R is some recursive subset of A . But $R \subset^* Y$, since $Y \subset_m A$. Thus $A_r^0 \subset^* X \cup Y$.

LEMMA 2. Let A and Y (with $Y \subset_{sm} A$) be given. Then there is an r.e. sequence $(S_r)_{r \in 2^{<\omega}}$ of disjoint sets with

$$(1) Y \cup \{S_a : r \leq a\} =^* Y \cup A_r^0$$

$$(2) \text{ If } r \notin \Delta \text{ then } S_r \text{ is finite and } S_r \setminus Y \neq \emptyset.$$

PROOF OF THE THEOREM. Let $\Delta_0 \in E_1^3(\Delta)$ and W the r.e. subset as in Lemma 1 constructed by means of Δ_0 . Further let $(S_r)_{r \in 2^{<\omega}}$ be the r.e. sequence as in Lemma 2 (by means of A and some Y with $Y \subset_{sm} A$).

Let X be the set

$$Y \cup \bigcup \{S_r : r \in W\}$$

X is an r.e. subset of A , since W and $(S_r)_{r \in 2^{<\omega}}$ are r.e. We show that $\Delta_0 = \Delta_X$.

Suppose $r \in \Delta_0$. Then $[r] \subset^* W$, by Lemma 1, (3). For all $a \in S_a \setminus Y$ is finite, by Lemma 2(1). If S_a is infinite then we have even $S_a \subseteq Y$. Thus $X \supset Y \cup A_r^0$, since $[r] \subset^* W$. From this we get that $X \cup T[r]$ is r.e., hence $r \in \Delta_X$.

Suppose $r \notin \Delta_0$. Then, by Corollary 2 there is a branch Γ with $r \in \Gamma$, $\Gamma \cap \Delta_0 = \emptyset$ and $\Gamma \setminus W$ is infinite. $\Gamma \cap \Delta_0 = \emptyset$ implies $\Gamma \cap \Delta = \emptyset$, since $\Delta \subseteq \Delta_0$. Thus for all $a \in \Gamma$, $S_a \setminus Y \neq \emptyset$. Hence $\bigcup \{S_a : a \in \Gamma \setminus W\} \cap (A \setminus Y)$ is infinite, disjoint with $X \cap (A \setminus Y)$ and by Lemma 2(1) is included (mod Fin) into $A_r^0 \setminus Y$. Thus $Y \cup A_r^0 \not\subset^* X$. Using the fact proved before Lemma 2 $X \cup T[r]$ is not r.e. This means $r \notin \Delta_X$.

PROOF OF LEMMA 1. The r.e. set W is constructed stepwise. Let $W_0 = \emptyset$.

Simultaneously with W_s we construct a function h defined for all $s \geq 0$ and r with $|r| \leq s$ into ω such that the following properties are satisfied:

- (1) $|r| < h(r, s + 1) < h(r * i, s + 1)$ (for all r with $|r| \leq s$)
- (2) $\max |\Psi_{r,s}| < h(r, s + 1)$
- (3) $[r, h(r, s + 1)] \cap W_s = \emptyset$
- (4) h is the smallest function satisfying (1), (2) and (3) (i.e. first $h(\langle \ \rangle, s + 1)$ becomes defined as small as possible, then $h(0, s + 1)$ and $h(1, s + 1)$ are defined and so on until $h(r, s + 1)$ with $|r| = s + 1$. Thus $h(\langle \ \rangle, 0) = 1$).

Let W_{s+1} be W_s united with

$$\{a : |a| < h(\langle \ \rangle, s + 1)\} \cup \{a : r * i \leq a \wedge h(r, s + 1) < |a| < h(r * i, s + 1)\}.$$

The set W is the union of all sets $W_s, s \geq 0$. First we see that $h(\cdot, s + 1)$ is well-defined for all $s \geq 0$, since W_s is a subset of $\{r \in 2^{< \omega} : |r| \leq \max_{|a|=s} |h(a, s)|\}$ and thus is a finite set.

The condition (3) always can be satisfied.

Further the set W_{s+1} is well-defined, since $|r * i| \leq h(r, s)$, by (1).

Suppose $r \notin \Delta$. Then for all $a \leq r, a \notin \Delta$, hence $\lim_s \Psi_{a,s}$ exists. If s_0 is so that $\Psi_{a,s_0} = \lim_s \Psi_{a,s}$, then for all $s \geq s_0, h(a, s) = h(a, s_0)$ by (4). Hence $\lim_s h(a, s)$ exists.

But property (3) insures that $[r, h(r, s_0)] \cap W = \emptyset$. Thus (4) in Lemma 1 is satisfied.

Suppose $r \in \Delta$. Then for almost all $a \in [r] \lim_s \Psi_{a,s} = \omega$. Let r_1, \dots, r_k be the smallest words respectively to \leq inside $[r]$ with $\lim_s \Psi_{r_i,s} = \omega$. (Observe that for every $a \in [r]$ there is an r_i such that $a \leq r_i$ or $r_i \leq a$.)

Let s_0 be such that for all a with $a \leq r_i, a \neq r_i$ for some $i, h(a, s_0) = h(a, s)$ for $s \geq s_0$ and M the maximum of all these $h(a, s_0)$'s.

Then every $c \in [r]$ with $M < |c|$ comes into W , since all $\lim_s h(r_i, s) = \omega$.

PROOF OF LEMMA 2. Let a_0, a_1, \dots be an enumeration of A and Y_s the set Y after s steps of its enumeration.

STEP 0. Define $S_{r,0} = \emptyset$ for all $r \in 2^{< \omega}$.

STEP $s + 1$.

- (1) If a_s belongs to Y_s then define all $S_{r,s+1} = S_{r,s}$.

- (2) If $a_s \notin Y_s$ then look for the r_s with $a_s \in A_{r_s}$. (Since all A_r are disjoint and their union is equal to A , there is exactly one such word.)

Let $r_1|r_2$ in the following mean that $r_1 \leq r_2$ or $r_2 \leq r_1$. Look if

$$(\exists r \in 2^{>\omega})(r|r_s \wedge \max |S_{a,s}| < a_s \text{ for all } a \text{ with} \\ a|r_s \wedge a <_1 r \wedge a_s < \max |S_{r,s}|). \tag{5}$$

If there is such an r it is unique, since $<_1$ is a linear ordering. Let r_0 be this word.

CASE (a). If there is an a with $a|r_s \wedge a <_1 r_0$ and $S_{a,s} \subseteq Y_s$, then put a_s into $Y_{a',s}$, where a' is the smallest such a .

CASE (b). If such an a does not exist, then put a_s into $S_{r_0,s+1}$. (All other $S_{r,s+1}$ are defined equal to $S_{r,s}$.)

If (5) does not hold, hence $\max |S_{r,s}| < a_s$ for all r with $r|r_s$, then let r_1 be the smallest word with $r_1|r_s$ such that $S_{r_1,s} \subseteq Y_s$. (Such a word always exists, since almost all sets $S_{r,s}$ are even empty.) Define $S_{r_1} = S_{r_1,s} \cup \{a_s\}$.

Let $S_r = \bigcup_{s \geq 0} S_{r,s}$. From the construction it follows that $(S_r)_{r \in 2^{<\omega}}$ is an r.e. sequence of disjoint r.e. sets (with $Y \cup \bigcup S_r = A$).

— If for some $S_r \setminus Y \neq \emptyset$ then S_r is finite.

Let $S_r \setminus Y \neq \emptyset$. Then for some s_0 we have $S_{r,s} \setminus Y \neq \emptyset$ for all $s \geq s_0$. Thus after step s_0 an element a from A comes to $S_{r,s+1}$ only if $a < \max |S_{r,s_0}|$. But there are only finitely many such elements.

— For all $r \ Y \cup A_r^0 = *Y \cup \{S_a : r \leq a\}$.

By the construction we know that an $x \in A_r$ comes into S_a then $a|r$. Thus elements from $A_r^0 \setminus Y$ if they are put into S_a and not $r \leq a$ then $a < r$. But such a 's are only finitely many and every S_a with $S_a \setminus Y \neq \emptyset$ is finite. Hence almost all $x \in A_r^0 \setminus Y$ comes to some S_a with $r \leq a$. This shows that \subseteq . For \supseteq first we see that for all $r \ A_r \setminus Y$ is finite. (If $r \notin \Delta$ then A_r is finite and if $r \in \Delta$ $A_r \subseteq A_r^0$ and $A_r^0 \setminus Y$ is finite.)

Thus for every r the set $\bigcup \{A_a \setminus Y : a < r\}$ is finite and other elements from $A \setminus Y$ do not come into S_a for $r \leq a$. Hence we have \supseteq .

— Let $r \notin \Delta$. Then $S_r \setminus Y \neq \emptyset$.

Suppose not and let r be the smallest word w.r.t. $>_1$ with $r \notin \Delta$, but

$S_r \setminus Y = \emptyset$. Since $r \notin \Delta$, A_r^0 is not recursive. Further, since $Y \subseteq_{sm} A$ $A_r^0 \setminus Y$ is not co-r.e. For all $a < r$, $a \notin \Delta$, and by the choice of r $S_a \setminus Y \neq \emptyset$. From above we have that these S_a 's are finite.

Let s_0 be such that for all $a < r$ $S_{a,s_0} = S_a$. Let s_1 be such that $s_0 \leq s_1$ and for $s \geq s_1$ a_s is greater than every $\max |S_a|$, $a < r$.

First we show that S_r must be infinite. From $S_r \setminus Y \neq \emptyset$ it follows that infinitely often $S_{r,s} \subseteq Y_s$. If not then for almost all s , $S_{r,s} \subseteq Y_s$. But then an a_s comes into $S_{r,s+1}$ only if $a_s \leq \max |S_{r,s+1}|$. Hence S_r is finite and so for some s' $S_{r,s'} = S_r$. This and $S_{r,s'} \setminus Y \neq \emptyset$ implies $S_r \setminus Y \neq \emptyset$, which contradicts the assumption.

Let $F(s)$ be the function equal to $\max |S_{r,s}|$. Then F is recursive and increasing and so the set $\{a_s : F(s) < a_s\}$ is recursive and includes (mod Fin) $A_r^0 \setminus Y$. If not, then at least one element of this set comes into S_r , by the construction.

Thus $\{a_s : F(s) < a_s\} \cap (A_r^0 \cup T[r])$ is recursive and includes (mod Fin) $A_r^0 \setminus Y$. But this contradicts that $A_r^0 \setminus Y$ is not co-r.e.

We see that this Theorem extends the Theorem 6 of LACHLAN (1968), since the isomorphism between both structures in the Theorem of course maps the principal ideals of $P_1(A)$ onto $E_{1,r}^3(\Delta)$. Further $E_{1,r}^3(\Delta)$, $\Delta \in E_1^{3-}$ are just all $\exists \forall \exists$ -Boolean algebras.

QUESTION. Since the ideal family $P_1(\cdot)$ is defined by lattice definable notions, the isomorphism between $P_1(A)$ and $P_1(B)$ is a necessary condition for hh-simple sets A and B to be automorphic. Is the isomorphism between both ideal families sufficient for the automorphism between both?

Minimal ideal families

From Theorem 1 it follows that $P_1(A)$ for all hh-simple sets is a lattice and that for many hh-simple sets the corresponding ideal families include also nonprincipal ideals. Thus e.g. the atomless hh-simple set constructed in Theorem 5 (LACHLAN 1968) is just so that $P_1^*(A) \cong E_1^3(\emptyset) = E_1^3$, hence includes all possible ideals.

The following Theorem treats the opposite situation:

THEOREM 2 (HERRMANN 1986). *For every $\exists \forall \exists$ -Boolean algebra \mathfrak{A} there is an hh-simple set A such that $L^*(A) \cong \mathfrak{A}$ and $P_1^*(A)$ includes only principal ideals.*

SKETCH OF THE PROOF. We have to show that for a given Δ' there is an $\Delta \in E_1^3$ such that $E_{1,r}^3(\Delta') \cong E_{1,r}^3(\Delta)$ and $E_{1,r}^3(\Delta) = E_1^3(\Delta)$. Since all $\exists\forall\exists$ -Boolean algebras are of the form $E_{1,r}^3(\Delta')$ for $\Delta' \in E_1^3$ and by Theorem 1, it is sufficient.

Let $\Delta \in E_1^3$ and $\text{Sp}(\Delta) = \{r: r * 0 \not\subseteq \Delta \wedge r * 1 \not\subseteq \Delta\}$.

(1) It is not difficult to see that for $\Delta_1, \Delta_2 \in E_1^3$

$$(\text{Sp}(\Delta_1), \leq) \cong (\text{Sp}(\Delta_2), \leq) \rightarrow E_{1,r}^3(\Delta_1) \cong E_{1,r}^3(\Delta_2). \tag{6}$$

(2) By $(\Delta_e)_{e \geq 0}$ we denote an Σ_3^0 -sequence consisting of all Σ_3^0 -ideals. (The existence of such a sequence is easy to prove.) Let $\text{Sp}_n(\Delta) = \{r \in \text{Sp}(\Delta): \text{card}(\{a < r: a \in \text{Sp}(\Delta)\}) = n\}$. We see that if for an $\Delta \in E_1^3$, every $e \geq 0$ and $r \in \text{Sp}_e(\Delta)$

$$[r] \subseteq \Delta_e \text{ or } [r] \cap \Delta_e \subseteq \Delta, \tag{7}$$

then

$$E_1^3(\Delta) = E_{1,r}^3(\Delta).$$

The combination of (1) and (2) gives the wanted ideal. For a given Δ' we construct (stepwise) Δ such that for Δ' and Δ (6) holds and Δ additionally satisfies (7).

Let r_0, r_1, \dots be an Σ_3^0 -sequence of the elements of Δ' and Δ'_{s+1} the smallest ideal including $\{r_0, \dots, r_s\}$ ($\Delta'_0 = \emptyset$).

Let $\Delta_0 = \emptyset$ and Δ_{s+1} the ideal such that the following conditions are satisfied

- $\Delta_s \subseteq \Delta_{s+1}$
- $(\text{Sp}(\Delta'_{s+1}), \leq) \cong (\text{Sp}(\Delta_{s+1}), \leq)$ (let p be an isomorphism)
- $|r| \leq |p(r)|, r \in \text{Sp}(\Delta_{s+1})$
- the elements of $\text{Sp}(\Delta_{s+1})$ are chosen minimal respectively to the following sequence of priority:

$$[e_{\langle \cdot \rangle}] \subseteq \Delta_{0,s}; |e_{\langle \cdot \rangle}|; [e_0] \subseteq \Delta_{1,s}; |e_0|; \dots [e_r] \subseteq \Delta_{r,s}; |e_r|; \dots,$$

where $e_r \in \text{Sp}_{|r|}(\Delta_{s+1})$ and $\Delta_{e,s}$ is an Σ_3^0 -enumeration of $(\Delta_e)_{e \geq 0}$. The set $\Delta = \bigcup_{s \geq 0} \Delta_s$ then is the ideal.

Number of orbits of hh-simple sets

By using Theorem 2 we can give an answer to a problem raised by SOARE (1978) about the number of orbits of hh-simple sets with isomorphic r.e. superset structures.

THEOREM 3 (HERRMANN 1986). *For every infinite $\exists\forall\exists$ -Boolean algebra \mathfrak{A} there is a sequence $(A_n)_{n \geq 0}$ of hh-simple sets with $L^*(A_n) \cong \mathfrak{A}$, but A_n and A_m ($n \neq m$) are not automorphic.*

SKETCH OF THE PROOF. Denote with $\oplus 2$, $I(\eta^+)$, $\square(\omega, \eta^+)$ the types of Boolean algebras: The B.A. of finite and cofinite subsets of ω , the countable atomless B.A., the B.A. which factorized by the ideal of atoms is the countable atomless algebra.

If \mathfrak{A} is a countable infinite Boolean algebra then there is an $a \in \mathfrak{A}$ such that $\{b \in \mathfrak{A} : b \leq a\}$ is isomorphic to $\oplus 2$, $I(\eta^+)$, or to $\square(\omega, \eta^+)$. Thus, if $L^*(A)$ is infinite then there is a recursive set R such that $L^*(A \cup R) \cong \oplus 2$, $I(\eta^+)$ or $\square(\omega, \eta^+)$. Thus it is sufficient to show the Theorem only for these three cases.

$\oplus 2$ was already shown in HERRMANN (1983).

For $I(\eta^+)$ we describe Σ_3^0 -ideals Δ_n , $n \geq 0$. The hh-simple sets A_n constructed by means of these are the wanted sets.

Let $\Delta_{\min,a}$ be the ideal from Theorem 2 with $E_{1,r}^3(\Delta_{\min,a}) \cong I(\eta^+)$. Denote by a_i the word $1 \dots 1$ (i -times) and b_i the word $0 \dots 0$ (i -times).

Δ_n are the sets

$$\{r : r = a_n * c, c \in \Delta_{\min,a} \text{ or } r = a_k * b_j * c; 0 \leq k < n, j \geq 0, c \in \Delta_{\min,a}\}.$$

For $\square(\omega, \eta^+)$ the Δ_n 's are defined exactly in the same way, but instead of $\Delta_{\min,a}$ we take the corresponding ideal from Theorem 2 for $\square(\omega, \eta^+)$.

Final remarks

The automorphism analysis of the hh-simple sets is an extensive topic for itself. There are still many open problems and questions. Of course the main problem is to find a necessary and sufficient condition for two hh-simple sets to be automorphic. The isomorphism type of the family

$P_1^*(A)$ for the hh-simple set A could be such a condition. Still open is the proof of the implication

$$L^*(A) \cong_{\Sigma_4^0} L^*(B) \wedge P_1^*(A) \cong P_1^*(B) \rightarrow A \cong_E B.$$

But to show this, since such an automorphism is not Σ_3^0 -presented in general, a generalization of Soare's Extension Theorem to non-effective constructions of array extensions would be necessary.

References

- HERRMANN, E., 1983, *Orbits of hyperhypersimple sets and the lattice of Σ_3^0 sets*, J. Symb. Logic 48, pp. 693–699.
- HERRMANN, E., 1986, *Automorphisms of the lattice of recursively enumerable sets and hyperhypersimple sets*, Seminarberichte der Sektion Mathematik der Humboldt-Universität, Berlin Nr. 86, pp. 69–108.
- LACHLAN, A.H., 1968, *On the lattice of recursively enumerable sets*, Trans. Amer. Math. Soc. 130, pp. 1–37.
- LERMAN, M., SHORE, R.A., SOARE, R.I., 1984, *R-maximal major subsets*, Isr. J. Math. 31, pp. 1–18.
- MAASS, W., 1984, *On the orbit of hyperhypersimple sets*, J. Symb. Logic 49, pp. 51–62.
- ROGERS, H., 1972, *Theory of recursive functions and effective computability* (Russian edition), Isdatelstwo "Mir", Moscow.
- SOARE, R.I., 1974, *Automorphisms of the lattice of recursively enumerable sets*, Part I: *Maximal sets*, Ann. Math. 100, pp. 80–100.
- SOARE, R.I., 1978, *Recursively enumerable sets and degrees*, Bull. Amer. Math. Soc. 84, pp. 1149–1181.

DEGREES OF FUNCTIONS WITH NO FIXED POINTS*

CARL G. JOCKUSCH, Jr.

Department of Mathematics, University of Illinois, Urbana, IL 61801, USA

Let W_e be the e th r.e. subset of ω ($=\{0, 1, 2, \dots\}$) under a standard enumeration. The fixed-point form of Kleene's recursion theorem asserts that for every recursive function $f: \omega \rightarrow \omega$ there exists an $e \in \omega$ with $W_e = W_{f(e)}$. This paper will discuss results about the (Turing) degrees of functions which are *fixed-point free* (FPF) in the sense that $W_e \neq W_{f(e)}$ for all $e \in \omega$. This is easily seen to coincide with the class of degrees of all functions g which are *diagonally non-recursive* (DNR) in the sense that $g(e) \neq \varphi_e(e)$ for all $e \in \omega$, where φ_e is the e th partial recursive function in a standard enumeration. Thus we are also studying the degrees of functions whose non-recursiveness is ensured by the diagonal method. A number of results in this paper were obtained as part of an effort to decide whether there are diagonally non-recursive functions of minimal degree (either among all non-zero degrees or among the degrees of diagonally non-recursive functions), but this question remains open. Nonetheless, it is hoped that the results of this paper will shed some light on the scope of the diagonal method.

It will be obvious from this paper that the author is indebted to many people for information and help. He would particularly like to thank Marat Arslanov for introducing him to the subject of this paper as well as for his important contributions to it.

To establish notation, let f, g, h, p, r be variables for total functions from ω into ω , and let other small Roman letters be variables ranging over ω . If α and β are partial functions, $\alpha(e) = \beta(e)$ means that either α and β are defined with the same value at e , or both are undefined at e .

* This research was supported by the National Science Foundation.

DEFINITION 1.

- (a) $\text{FPF} = \{f : (\forall e)[W_e \neq W_{f(e)}]\}$,
 (b) $\text{DNR} = \{g : (\forall e)[g(e) \neq \varphi_e(e)]\}$.

The next result, although very easy, is important in that it establishes a close connection between FPF and DNR. Actually, DNR is usually easier to work with than FPF and we will generally work with DNR rather than FPF.

PROPOSITION 1 (JOCKUSCH *et al.* 1990: Lemma 4.1). *The degrees of functions in FPF coincide with the degrees of functions in DNR.*

PROOF. First note by easy coding arguments that both classes of degrees mentioned in Proposition 1 are closed upwards. (For FPF use that, given i , one may effectively find $j \neq i$ with $W_j = W_i$. For DNR use that K is infinite and r.e., where $K = \{e : \varphi_e(e) \downarrow\}$.)

Now given $f \in \text{FPF}$ we must find $g \in \text{DNR}$ with g recursive in f . Let h be a recursive function with $W_{h(e)} = W_{\varphi_e(e)}$ for all $e \in \omega$, and let $g(e) = f(h(e))$. Then, for all e , $W_{f(h(e))} \neq W_{h(e)} = W_{\varphi_e(e)}$, so $g(e) \neq \varphi_e(e)$ as required.

Finally given $g \in \text{DNR}$ we must find $f \in \text{FPF}$ with f recursive in g . Let h be a recursive function with $\varphi_{h(e)}(x) \in W_e$ for all e with $W_e \neq \emptyset$ and all x , and let f be a function recursive in g with $W_{f(e)} = \{g(h(e))\}$ for all e . Assume that $W_e = W_{f(e)}$ in order to obtain a contradiction. Then $W_e \neq \emptyset$ so $\varphi_{h(e)}(h(e)) \in W_e = W_{f(e)}$. It follows that $\varphi_{h(e)}(h(e)) = g(h(e))$, which contradicts the hypothesis that $g \in \text{DNR}$. \square

Note that Proposition 1 implies the recursion theorem. Indeed the second paragraph of its proof is essentially the same as the proof of the recursion theorem.

It is clear that every degree $\geq \mathbf{0}'$ (the degree of the halting problem) contains a function in FPF. The existence of a degree $\mathbf{a} < \mathbf{0}'$ in FPF was first proved by M. Arslanov by a complicated direct construction and later using the low basis theorem (JOCKUSCH and SOARE 1972) in connection with the notion of effective simplicity (see ARSLANOV 1981). The following proposition illustrates another method of using results about Π_1^0 classes to obtain information about FPF.

COROLLARY 1 (ARSLANOV 1981). (a) *There is a degree \mathbf{a} in FPF with $\mathbf{a}' = \mathbf{0}'$, where \mathbf{a}' denotes the jump of \mathbf{a} .* (b) *There is a degree \mathbf{a} in FPF such that*

every function of degree $\leq \mathbf{a}$ is majorized by a recursive function. (c) There are two degrees in FPF which have infimum $\mathbf{0}$. (d) There is a degree $\leq \mathbf{0}'$ in FPF which does not have any non-zero r.e. predecessor.

PROOF. Let DNR_2 be the class of all functions $g \in \text{DNR}$ with $g(e) < 2$ for all $e \in \omega$. Then DNR_2 is a non-empty, recursively bounded Π_1^0 class of functions. By the low basis theorem (JOCKUSCH and SOARE 1972: Theorem 2.1), DNR_2 has an element of low degree (i.e. of degree \mathbf{a} with $\mathbf{a}' = \mathbf{0}'$). This proves (a) and (b), (c) and (d) follow similarly from Theorem 2.4, Corollary 2.9 and Corollary 2.11, respectively of JOCKUSCH and SOARE (1972).

In spite of Corollary 2, it is possible to obtain results which show that important subclasses of DNR contain only degrees which are "far from $\mathbf{0}$ ". The most fundamental of these results is known as the Arslanov completeness criterion.

THEOREM 1 (ARSLANOV 1981: Theorem 1). *An r.e. set A has degree $\mathbf{0}'$ if and only if some function in FPF is recursive in A .*

Theorem 1 has been applied to many basic constructions in recursion theory, such as the Post simple set construction and the maximal set constructions of Friedberg and Yates, to show that these constructions automatically yield complete r.e. sets and thus that Theorem 1 implies in a sense that Post's problem cannot be solved by the diagonal method (see e.g. ARSLANOV *et al.* 1977).

The following extension of the Arslanov completeness criterion was obtained by JOCKUSCH *et al.* (1990) following a previous attempt by ARSLANOV (1981).

THEOREM 2. *Let A be a finite Boolean combination of r.e. sets. Then A has degree $\mathbf{0}'$ if and only if some function in FPF is recursive in A .*

The proof of Theorem 2 is inherently less uniform than that of Theorem 1. Specifically, Theorem 1 holds uniformly in the sense that there is an effective procedure which, given an index of an r.e. set A and given an index for computing a function in FPF from oracle A , produces an index for computing K from A . On the other hand, it is shown by JOCKUSCH *et al.* (1990: Theorem 6.4) that the corresponding uniformity fails even for differences of r.e. sets. Thus there is no effective procedure which, given a , b , and c such that $\Phi_c(W_a - W_b)$ is a function in FPF,

produces a number e with $\Phi_e(W_a - W_b) = K$. (Here we write Φ_e for the e th reduction procedure in a standard enumeration.)

Theorem 2 is optimal with respect to the Ershov difference hierarchy (ERSHOV 1968/1970, EPSTEIN *et al.* 1981). The finite Boolean combinations of r.e. sets are exactly the sets which occur at the finite levels of this hierarchy, and the sets reducible to K by truth-tables are the sets which occur at level ω of this hierarchy. An analysis of the proof of the low basis theorem (JOCKUSCH and SOARE 1972: Theorem 2.1) shows that every non-empty Π_1^0 class of sets has an element A such that A (in fact A') occurs by level ω of the difference hierarchy. Applying this to DNR_2 , one obtains a degree \mathbf{a} such that $\mathbf{a}' = \mathbf{0}'$, $\mathbf{a} \in \text{FPF}$, and every set of degree \mathbf{a} occurs by level ω of the difference hierarchy. Nonetheless, it is possible to extend Theorem 2 by considering a broader hierarchy, the REA hierarchy, which was originally introduced by ARSLANOV (1982). A set A is called 1 -REA if it is r.e. and $(n+1)$ -REA if it is r.e. in some n -REA set B with B recursive in A . Every finite Boolean combination of r.e. sets is n -REA for some n , but there are 2 -REA sets of degree $\leq \mathbf{0}'$ which are not finite Boolean combinations of r.e. sets (JOCKUSCH and SHORE 1984: Theorem 1.6). Note also that n -REA sets need not be recursive in $\mathbf{0}'$ for $n > 1$, and indeed $0^{(n)}$ is an n -REA set.

THEOREM 3 (JOCKUSCH *et al.* 1990: Theorem 5.1). *If A is n -REA for any n , then A has degree $\geq \mathbf{0}'$ if and only if some function in FPF is recursive in A .*

Recently Kučera has obtained a surprising result on FPF which contrasts strongly with Corollary 1.

THEOREM 4 (KUČERA 1986: Theorem 1). *For any degree $\mathbf{a} \leq \mathbf{0}'$ in FPF, there is a non-zero r.e. degree $\mathbf{b} \leq \mathbf{a}$. In fact, \mathbf{b} may be chosen to be promptly simple (KUČERA 1986: Remark 2).*

One remarkable aspect of Theorem 4 is that it is proved without any use of the priority method. Thus in combination with Corollary 1(a) (which is also proved without the priority method), Theorem 4 yields a priority-free solution to Post's problem! See (KUČERA 1986, 1988) for a discussion of further aspects of this method, and see (KUČERA 1987) for an extension to infinite injury arguments such as the construction of an incomplete high r.e. degree.

Let DNR_k be the class of all functions $g \in \text{DNR}$ with $g(e) < k$ for all e . We now show that no functions of minimal degree are in DNR_k for any k .

PROPOSITION 2 (JOCKUSCH and SOARE 1972; SOLOVAY, unpublished). *The degrees of the functions in DNR_2 coincide with the degrees of complete extensions of Peano arithmetic (PA).*

PROOF. First, let g be any function in DNR_2 . Then g is the characteristic function of a set which separates $\{e : \varphi_e(e) = 0\}$ and $\{e : \varphi_e(e) = 1\}$ and it is well known that these two sets are effectively inseparable. Then by (JOCKUSCH and SOARE 1972: Proposition 6.1), there is a complete extension of Peano arithmetic which is recursive in g . For the other direction, let T be any complete extension of PA. By a result of SCOTT (1962), every non-empty recursively bounded Π_1^0 class (in particular DNR_2) has an element recursive in T .

The proof of Proposition 2 will now be complete once we show that the class of degrees of functions in DNR_2 and the class of degrees of complete extensions of Peano arithmetic are each closed upwards in the degrees, i.e. for any degree \mathbf{a} in the class, all degrees $\mathbf{b} \geq \mathbf{a}$ are also in the class. For DNR_2 the upward closure result is very easy. The upward closure of the degrees of complete extensions of Peano arithmetic is due to SOLOVAY (unpublished) and answers a question raised by JOCKUSCH and SOARE (1972). Let T^* be a complete extension of PA and suppose that T^* is recursive in C . We must construct a complete extension T of PA of the same degree as C . Let $\gamma_0, \gamma_1, \dots$ be a recursive enumeration of all sentences in the language of PA. We obtain T as the union of an ascending chain of theories T_0, T_1, \dots . Let T_0 be PA. Given T_{2n} , use T^* to T^* -effectively compute a formula $\beta_n \in \{\gamma_n, \neg\gamma_n\}$ such that $T_{2n} \cup \{\beta_n\}$ is consistent if T_{2n} is, and let T_{2n+1} be $T_{2n} \cup \{\beta_n\}$. Given T_{2n+1} , effectively compute a formula ψ_n such that both ψ_n and $\neg\psi_n$ are consistent with T_{2n+1} if T_{2n+1} is consistent. Let T_{2n+2} be $T_{2n+1} \cup \{\psi_n\}$ if $n \in C$, and otherwise let T_{2n+2} be $T_{2n+1} \cup \{\neg\psi_n\}$. Clearly $T = \bigcup_n T_n$ is a complete extension of PA and has the same degree as C . \square

For more information on the degrees of the complete extensions of PA, see, for instance, SCOTT (1962), JOCKUSCH and SOARE (1972), SIMPSON (1977) and KUČERA (1985, 1986, 1987, 1988). The next result shows, in conjunction with Proposition 2, that these degrees coincide with the degrees of DNR functions bounded by a constant.

THEOREM 5. *For each $k \geq 2$, the degrees of functions in DNR_k coincide with the degrees of functions in DNR_2 .*

PROOF (FRIEDBERG, R., unpublished). Since it is easily seen that the degrees of functions in DNR_k are closed upwards, it suffices to show that given any function g in DNR_k there is a function f recursive in g which is in DNR_2 and is recursive in g . This is done by induction on k , and in fact we show that if $g \in \text{DNR}_{k^2}$, then there is a function h recursive in g with $h \in \text{DNR}_k$. Identify each natural number with the set of all smaller natural numbers. We first obtain two functions $g_1(a, b)$ and $g_2(a, b)$, each recursive in g and taking values less than k , such that for all a and b

$$g_1(a, b) \neq \varphi_a(a) \quad \text{or} \quad g_2(a, b) \neq \varphi_b(b). \quad (1)$$

To define g_1 and g_2 , fix a recursive pairing function $\langle -, - \rangle$ from ω^2 to ω which also maps $k \times k$ to k . Given a and b , effectively compute c so that $\varphi_c(c) = \langle \varphi_a(a), \varphi_b(b) \rangle$. Then $g_1(a, b)$ and $g_2(a, b)$ are determined by the equation $g(c) = \langle g_1(a, b), g_2(a, b) \rangle$. Since $g(c) \neq \varphi_c(c)$, equation (1) above holds.

We now consider two cases.

CASE 1. For every a there exists b with $g_2(a, b) = \varphi_b(b)$. Given a , find such a b by g -effective search, and set $h(a) = g_1(a, b)$.

CASE 2. There is an a with $g_2(a, b) \neq \varphi_b(b)$ for all b . Fix such an a , and set $h(b) = g_2(a, b)$ for all b .

In both cases, it is immediate that h is recursive in g and is in DNR_k , as required. \square

COROLLARY 2. *No function in DNR_k for any k is of minimal degree. In fact, any countable partially ordered set can be embedded in the ordering of the degrees below any function in any DNR_k .*

PROOF. Corollary 2 is immediate from Proposition 2, Theorem 3, and (JOCKUSCH and SOARE 1972, Corollary 4.4), which asserts that any countable, partially ordered set can be embedded in the degrees below the degree of any complete extension of PA. \square

COROLLARY 3. *No function in FPF bounded by a constant is of minimal degree.*

PROOF. The proof of Proposition 1 shows that given f in FPF bounded by k , there exists g recursive in f with $g \in \text{DNR}_k$. Apply Corollary 2.

It would be pleasant if Theorem 5 could be extended to show that the degrees of all functions in FPF coincide with the degrees of complete extensions of PA. However, this result fails, as Kučera has pointed out. The degrees of complete extensions of PA have measure 0 by (JOCKUSCH and SOARE 1972, Corollary 5.4), and yet the degrees of functions in FPF have measure 1 by (KUČERA 1985, Corollary 2). By a more refined version of this argument (unpublished) Kučera has in fact shown that there is a degree below $\mathbf{0}'$ which is in FPF and yet contains no complete extension of PA. One might still hope that functions in DNR which are "sufficiently small" in some sense (weaker than being bounded by a constant) might have to bound complete extensions of PA. This is certainly not so for functions in DNR bounded by recursive functions, since the degrees of such functions have measure 1. Second, given any recursive function p not bounded by a constant and any function $g \in$ DNR bounded by a recursive function, there is a Gödel-numbering of the partial recursive functions and a function g' in DNR (with respect to the given Gödel-numbering) which is bounded by p and recursive in g . Thus almost every degree contains a DNR function (with respect to the given Gödel-numbering, which depends on p) which is bounded by p . The next result gives bounds which work independently of the Gödel-numbering.

PROPOSITION 3 (KURTZ, S., unpublished). *Let $p(n)$ be a recursive function with $p(n) \geq 2$ for all n such that $\sum_n (1/p(n))$ is convergent. Then for almost every set A , there is a function g^A recursive in A which is in DNR with $g^A(n) < p(n)$ for all n .*

PROOF. First let $h^A(n) = \sum \{2^k : k \in A \text{ \& } k < n\}$. Clearly $h^A(n) < 2^n$ for all n , and all such values of $h^A(n)$ are equally likely in the usual measure on 2^ω . Let $f^A(n)$ be the remainder when $p(n)$ is divided into $h^A(n)$. Thus $f^A(n) < p(n)$ for all n , and for any $i < p(n)$, the probability that $f^A(n) = i$ is at most $1/p(n) + 1/2^n$. Given a real number $\epsilon > 0$, choose n_0 sufficiently large that

$$\sum_{n=n_0}^{\infty} \left(\frac{1}{p(n)} + \frac{1}{2^n} \right) < \epsilon.$$

Define $g^A(n)$ to be $f^A(n)$ for $n \geq n_0$ and for $n < n_0$ let $g^A(n) = 1$ if $\varphi_n(n) = 0$, and otherwise let $g^A(n) = 0$. Then an elementary calculation shows that $g^A \in$ DNR with probability at least $1 - \epsilon$. \square

As the reader probably noticed, the proof of Theorem 5 was quite non-uniform in the sense that the division between Cases 1 and 2 was not effective, and also the number a in Case 2 is not found effectively. We now show that this lack of uniformity is essential.

THEOREM 6. *There is no reduction procedure Φ and number k such that $\Phi(g) \in \text{DNR}_k$ for all $g \in \text{DNR}_{k+1}$.*

PROOF. Suppose there were such a reduction procedure Φ and number k . We now apply the recursion theorem to obtain a number e where Φ goes wrong, i.e. $\Phi(g; e) = \varphi_e(e)$ for some $g \in \text{DNR}_{k+1}$. A *string* is a finite sequence of natural numbers. We use the letters σ , τ , and μ for strings. A string σ is in DNR if $\sigma(e) \neq \varphi_e(e)$ for all e in the domain of σ . We define DNR_{k+1} for strings in the obvious way. Since $\Phi(g)$ is total for all $g \in \text{DNR}_{k+1}$, by compactness there is a number t such that $\Phi(\sigma; e)$ is defined for all strings in DNR_{k+1} of length t . Such a number t may be found by effective search, specifically by looking for a stage s and a number t such that for every string σ of length t bounded on all arguments by $k+1$, either it is clear within s steps of computation that $\sigma \notin \text{DNR}$ or that $\Phi(\sigma; e)$ is convergent with value $< k$. We would now like to set $\varphi_e(e) = \Phi(\sigma; e)$, where σ is chosen to be in DNR_{k+1} , but it is not clear how to do this because there is no obvious way to obtain such a σ effectively. The solution is to show that there are so many σ 's giving the same value of $\Phi(\sigma; e)$ that we know one of them must be in DNR_{k+1} . In the following definition we fix s and t as above. Let D be the set of numbers z less than t such that $\varphi_z(z)$ is defined in at most s steps. Let T be the set of strings τ of length t with $\tau(z) < k+1$ for all $z < t$ and $\tau(z) \neq \varphi_z(z)$ for all $z \in D$. By hypothesis, $\Phi(\tau; e)$ is defined with value $< k$ for all strings $\tau \in T$. If $S \subseteq T$, call S *large* if for every string τ of length t with $\tau(z) < k+1$ for all $z < t$ there is a string $\sigma \in S$ with $\sigma(z) \neq \tau(z)$ for all $z < t$ with $z \notin D$. Note that any large set of strings must have an element in DNR_{k+1} . Also $T = \cup_{i < k} S_i$, where, for $i < k$, S_i is the set of strings $\sigma \in T$ with $\Phi(\sigma; e) = i$. We claim that S_i is large for at least one $i < k$. Once this is established, the argument is completed by setting $\varphi_e(e) = i$, for such an i , and clearly such an i can be found effectively if it exists. Suppose now for a contradiction that no S_i is large, and, for $i < k$, let τ_i be a string which witnesses that S_i is not large. Thus $\tau_i(z) < k+1$ for all $z < t$, and for all strings $\sigma \in S_i$, there exists $z < t$ with $z \notin D$ and $\sigma(z) = \tau_i(z)$. Now choose $\tau \in T$ so that, for all $z < t$, $\tau(z) < k+1$ and, if $z \notin D$, $\tau(z) \neq \tau_i(z)$ for all $i < k$. (This is possible because for $z < t$ with

$z \notin D$, there are $k + 1$ possible values for $\tau_i(z)$ and at most k are ruled out by equalling $\tau_i(z)$ for some $i < k$.) Clearly $\tau \notin S_i$ for all $i < k$, so we have a contradiction to the fact noted above that $T = \bigcup_{i < k} S_i$. This shows that at least one of the sets S_i is large and, as indicated above, the proof is completed by setting $\varphi_e(e) = i$ for some such i . \square

COROLLARY (to the proof). *For each k there exists $f \in DNR_{k+1}$ such that there is no $g \in DNR_k$ with $g \leq_{\text{tot}} f$, where $g \leq_{\text{tot}} f$ means there exists e with $\Phi_e(h)$ total for all functions h and $\Phi_e(f) = g$.*

The corollary is proved by a Kleene–Post style argument, with the construction of the theorem used to take care of each e with $\Phi_e(h)$ total for all functions h .

We now consider the connection between diagonal non-recursiveness and effective immunity. Recall that an infinite set A is called *effectively immune* if there is a recursive function p such that

$$(\forall e)[W_e \subseteq A \rightarrow |W_e| < p(e)]$$

MARTIN (1966) proved that every co-r.e. effectively immune set has degree $\mathbf{0}'$. Subsequently ARSLANOV, NADIROV, and SOLOV'EV (1977) deduced this result from Theorem 1 by noting that if A is effectively immune (say via p), then there is a function f recursive in A with $f \in \text{FPF}$, i.e. $W_{f(e)}$ consists of the first $p(e)$ elements of A . Thus, since FPF is upward closed as a set of degrees, every degree which contains an effectively immune set also contains a function in FPF. We now prove the converse of this.

THEOREM 7. *Every degree in FPF contains an effectively immune set.*

PROOF. By Proposition 1 and the upward closure of the degrees of effectively immune sets (JOCKUSCH 1973), it suffices to show that if $g \in DNR$, there is an effectively immune set A recursive in g . We show first that there is a function h recursive in g which has a strengthened form of the DNR property. Let

$$\text{SDNR} = \{h: (\forall e)(\forall u \leq e)[h(e) \neq \varphi_u(u)]\}$$

The idea of this definition is that functions in SDNR can diagonalize arbitrary effectively given finite sets of computations, as opposed to single computations for functions in DNR.

LEMMA. *For any function $g \in \text{DNR}$, there exists $h \in \text{SDNR}$ which is recursive in g .*

PROOF. As usual, let $(t)_e$ denote the $e + 1$ st term of the finite sequence of natural numbers coded by t , if this sequence has at least $e + 1$ terms. Let r be a recursive function such that $\varphi_{r(u)}(r(u)) = (\varphi_u(u))_u$ for all u . Let h be a function recursive in g such that, for all e and all $u \leq e$, $(h(e))_u = g(r(u))$, where g is a given function in DNR. To see that h is in SDNR, note that if $u \leq e$, then $(h(e))_u = g(r(u)) \neq \varphi_{r(u)}(r(u)) = (\varphi_u(u))_u$, so $h(e) \neq \varphi_u(u)$. This proves the lemma.

To complete the proof of the theorem, given $h \in \text{SDNR}$, we must construct an effectively immune set A recursive in h . We choose the elements of A in natural order by induction. Let these elements be a_0, a_1, \dots in order of size. To choose a_0 , let i_0 be an effectively chosen index so that $\varphi_{i_0}(i_0) \in W_0$ if $W_0 \neq \emptyset$. Let $a_0 = h(i_0)$. Assume inductively that a_0, \dots, a_n and indices i_0, \dots, i_n have been defined. From these compute an index i_{n+1} such that $\varphi_{i_{n+1}}(i_{n+1}) \in W_{n+1} - \{a_0, a_1, \dots, a_n\}$ if $W_{n+1} - \{a_0, a_1, \dots, a_n\} \neq \emptyset$. For each $u \leq a_n$ let e_u be a number e with $\varphi_e(e) = u$. Let m be the largest of the numbers $e_0, \dots, e_{a_n}, i_0, \dots, i_{n+1}$, and set $a_{n+1} = h(m)$. Thus a_{n+1} is chosen so that $a_{n+1} > a_n$ and $a_{n+1} \neq \varphi_{i_k}(i_k)$ for all $k \leq n + 1$.

Let $A = \{a_n : n \in \omega\}$. To show that A is effectively immune, it suffices to show that if $W_e \subseteq A$, then $W_e \subseteq \{a_i : i < e\}$. Assume this fails for $e = n + 1$. Then $W_{n+1} - \{a_i : i \leq n\} \neq \emptyset$, so $\varphi_{i_{n+1}}(i_{n+1}) \in W_{n+1} - \{a_i : i \leq n\} \neq \emptyset$. By construction, $a_{m+1} \neq \varphi_{i_{n+1}}(i_{n+1})$ for all $m \geq n$, so $\varphi_{i_{n+1}}(i_{n+1}) \notin A$. Since $\varphi_{i_{n+1}}(i_{n+1}) \in W_e$, the hypothesis that $W_e \subseteq A$ has been contradicted, and the proof of Theorem 7 is complete.

As mentioned at the beginning of this paper, we have not been able to decide whether there exist minimal degrees which contain functions in DNR. We do not even have a firm conjecture. However, the necessity of using non-uniform methods in a number of results mentioned in this paper suggests that such methods will also be required to decide whether there is such a minimal degree. It does seem highly unlikely that there is a single reduction procedure Φ with $\Phi(g)$ total for all $g \in \text{DNR}$ and $\phi <_T \Phi(g) <_T g$ for all $g \in \text{DNR}$. The proof of Corollary 2 does yield such a Φ for all $g \in \text{DNR}_2$, but it is not known whether there is such a Φ even for all $g \in \text{DNR}_3$. It is also not known whether there is a recursively bounded function in DNR which has minimal degree, nor whether every function in DNR is Turing equivalent to some recursively bounded function in DNR. However, a straightforward analysis of the proof of

Theorem 7 shows that the degrees of the recursively bounded DNR functions coincide with the degrees of the effectively immune, non-hyperimmune sets.

References

- AMBOS-SPIES, K., JOCKUSCH, JR., C.G., SHORE, R.A. and SOARE, R.I., 1984, *An algebraic decomposition of the recursively enumerable degrees and the coincidence of several degree classes with the promptly simple degrees*, Trans. Amer. Math. Soc. 281, pp. 109–128.
- ARSLANOV, M.M., 1981, *On some generalizations of a fixed point theorem*, Izv. Vyss. Uchebn. Zaved. Mat. 25(5), pp. 9–16 (in Russian); Sov. Math. (Iz. VUZ) 25 (5) (1981), pp. 1–10 (English translation).
- ARSLANOV, M.M., 1982, *On a hierarchy of the degrees of unsolvability*, Veroyatn. Metod. i Kibern. 18, pp. 10–17 (in Russian).
- ARSLANOV, M.M., NADIROV, R.F. and SOLOV'EV, V.D., 1977, *Completeness criteria for recursively enumerable sets and some general theorems on fixed points*, Izv. Vyss. Uchebn. Zaved. Mat. 179 (4), Sov. Math. (Iz. VUZ) 21 (4) (1977), pp. 1–4 (English translation).
- EPSTEIN, R.L., HAAS, R. AND KRAMER, R., 1981, *Hierarchies of sets and degrees below $\mathbf{0}'$* , in: M. Lerman, R.A. Shore and R.I. Soare, eds., Logic Year 1979–80: University of Connecticut, Lecture Notes in Mathematics, No. 859 (Springer, Berlin, Heidelberg, Tokyo, New York).
- ERSHOV, Y.L., 1968, *A hierarchy of sets*, Parts I, II, and III, Algebra i Logika 7, pp. 15–47 and 9 (1970), pp. 20–31 (Russian); English translation in Algebra and Logic 7 (1968), pp. 24–43 and pp. 212–232, and 9 (1970), pp. 20–31.
- JOCKUSCH, JR., C.G., 1973, *Upward closure and cohesive degrees*, Israel J. Math. 15, pp. 332–335.
- JOCKUSCH, JR. C.G. and SHORE, R.A., 1984, *Pseudo jump operators II: Transfinite iterations, hierarchies, and minimal covers*, J. Symbolic Logic 49, pp. 1205–1236.
- JOCKUSCH, JR. C.G., and SOARE, R.I., 1972, Π_1^0 -classes and degrees of theories, Trans. Amer. Math. Soc. 173, pp. 33–56.
- JOCKUSCH, JR., C.G., LERMAN, M., SOARE, R.I. and SOLOVAY, R.M., 1990, *Recursively enumerable sets modulo iterated jumps and Arslanov's completeness criterion*, to appear in J. Symbolic Logic.
- KUČERA, A., 1985, *Measure, Π_1^0 -classes and complete extensions of PA*, in: H.D. Ebbinghaus, G.H. Müller and G.E. Sacks, eds., Recursion Theory Week, Lecture Notes in Mathematics, No. 1141 (Springer, Berlin, Heidelberg, New York, Tokyo).
- KUČERA, A., 1986, *An alternative priority-free solution to Post's problem*, in: J. Gruska, B. Rován and J. Wiederman, eds., Twelfth Symposium held in Bratislava, Czechoslovakia, August 25–29, 1986, Lecture Notes in Computer Science, No. 233 (Springer, Berlin, Heidelberg, New York, Tokyo).
- KUČERA, A., 1987, *On the use of diagonally nonrecursive functions*, to appear in Logic Colloquium '87.
- KUČERA, A., 1988, *On the role of $\mathbf{0}'$ in recursion theory*, in: F.R. Drake and J. Truss, eds., Logic Colloquium '86 (North Holland, Amsterdam), pp. 133–141.
- MARTIN, D.A., 1966, *Completeness, the recursion theorem, and effectively simple sets*, Proc. Amer. Math. Soc. 17, pp. 838–842.
- SCOTT, D., 1962, *Algebras of sets binumerable in complete extensions of arithmetic*, in: Proc. Symp. Pure Math. 5 (American Mathematical Society, Providence, R.I.).
- SIMPSON, S.G., 1977, *Degrees of unsolvability: A survey of results*, in: K.J. Barwise, ed., Handbook of Mathematical Logic (North-Holland, Amsterdam, New York, Oxford).

This Page Intentionally Left Blank

4

Set Theory

This Page Intentionally Left Blank

FREE SETS FOR COMMUTATIVE FAMILIES OF FUNCTIONS

URI ABRAHAM

Ben Gurion University of the Negev, Department of Mathematics, Be'er Sheva, Israel

1. Introduction

Some of you are experts in fields outside mathematics and so I would like to devote part of my lecture to a description of some general aspects of set theory. Only the second part of my lecture is technical and deals with combinatorial problems the solution of which requires the methods developed by K. Gödel and P. Cohen and others for getting consistency results. In the first part I would like to explain what these methods for consistency results are capable of achieving and what in my opinion could be their impact on the field of set theory.

One of the basic notions developed by CANTOR (1845–1918) is that of a “power” or “cardinality”. Two sets A and B are said to *have the same cardinality* if and only if there is some correspondence between the elements of A and the elements of B which associates exactly one element of B to each element of A , and exactly one element of A to every element of B . For finite sets this is a familiar notion connected with the idea of counting: two finite sets have the same cardinality if and only if they have the same number of elements. But Cantor boldly applied this notion to infinite sets as well and thus provided the basis for extending the notion of “number” to the infinite.

An infinite set is called *denumerable* if and only if it has the same cardinality as the set of all integers. That is there is an enumeration $a_n, n = 1, 2, \dots$, of all elements of the set. For example the set of all even integers is denumerable because the correspondence $n \leftrightarrow 2n$ shows that the set of even integers has the same cardinality as the set of all integers. Cantor proved that the set of all rational numbers (the fractions of integers) is denumerable. And one of the earliest questions Cantor

asked was whether the set \mathfrak{R} of all real numbers is denumerable. Intuitively, it should not be denumerable because we feel that the continuous line (which is the geometric counterpart of the set of real numbers) is much richer than the set of integers. However, it took Cantor years to prove that \mathfrak{R} is not denumerable¹: There is no correspondence between the integers and the real numbers which enumerates all the reals.

Then Cantor looked at other subsets of \mathfrak{R} . He proved that any infinite closed subset of \mathfrak{R} is either denumerable or has the same cardinality as \mathfrak{R} . Several times Cantor believed that this result can be generalized to all subsets of \mathfrak{R} : that *any* infinite set of reals is either denumerable or has the same cardinality as the set of all real numbers. This statement is called the *continuum hypothesis*. In fact, at the end of his paper where the result about closed sets was proved, Cantor expressed the firm belief that in future publications he will prove the continuum hypothesis (see DAUBEN 1979: p. 118).

Today we know that it was impossible for Cantor to realize this dream, not because he was not a good enough mathematician but because it is a theorem that there is no proof of the continuum hypothesis or its negation (using the methods and intuition that were available in Cantor's time—or that are available today). In order to prove this so-called *independence* result we need first to formalize the methods of mathematics by which we think to prove facts about sets (that is, we have to give an axiomatic basis to set theory), and then we have to prove that those axioms do not derive the continuum hypothesis nor its negation. This is the consequence of two wonderful works by GÖDEL (1938) and by COHEN (1966) who showed the independence of the continuum hypothesis from the accepted axioms of set theory (the so-called ZF axioms—called after E. Zermelo and A. Fraenkel).

We can interpret Gödel's result as saying the following: if it is possible to imagine a world of all objects needed by the mathematicians (a universe of set theory) then it is possible to assemble only those objects which are definable in some intrinsic way. The resulting collection thus assembled (called the constructible sets) satisfies all the axioms of set theory and, moreover, form a universe of set theory in which the continuum hypothesis is true.

¹ See DAUBEN (1979: p. 50) (a citation from a letter of Cantor to Dedekind in 1873). Cantor proved this theorem by the method of shrinking closed intervals. I find it interesting to note that a theorem now taught to first year students was so hard to come by to Cantor himself. This shows how careful we should be in evaluating past achievements.

P. Cohen's work seems to look beyond: if it is possible to imagine a universe of set theory then it is possible to build a larger universe which is obtained by adding new sets. Such larger universes, called *generic extensions*, can be constructed to satisfy the negation of the continuum hypothesis.

Since both the collection of constructible sets and the generic extensions satisfy all the axioms of set theory (and thus each one contains all objects needed by the mathematicians) the continuum hypothesis or its negation cannot be derived from these axioms. Thus Cantor's continuum problem has no answer within the frame of established mathematics. This situation might leave us with some sense of frustration: it is in our nature to seek absolute and final answers to natural questions, and here is the major question in set theory which in principle seems to be unanswerable. Gödel [in section 3 of "What is Cantor's continuum problem?", see BENACERRAF and PUTNAM (1964)] wrote: "... a proof of the undecidability of Cantor's conjecture from the accepted axioms of set theory (in contradistinction, e.g., to the proof of the transcendency of π) would by no means solve the problem. For if the meaning of the primitive terms of set theory ... are accepted as sound, it follows that the set-theoretical concepts and theorems describe some well-determined reality, in which Cantor's conjecture must be either true or false. Hence its undecidability from the axioms being assumed today can only mean that these axioms do not contain a complete description of that reality." And P. Cohen himself wrote at the end of his book (see COHEN 1966, pp. 150–151): "One can feel that our intuition about sets is inexhaustible and that eventually an intuitively clear axiom will be presented which decides the continuum hypothesis." And then he adds: 'A point of view which the author feels may eventually come to be accepted is that the continuum hypothesis is *obviously* false.'

Now, today there does not seem to be in sight any axiom or intuition which might decide the truth of the continuum hypothesis — and I am not worried by this situation. There are many interesting results which rely on the continuum hypothesis and many other results which are meaningful in the presence of the negation of the continuum hypothesis; what would happen to them if the continuum hypothesis is decided in one way or the other? For example, assume that some intuition in favor of the negation of the continuum hypothesis is accepted, does this mean that all those constructions which use the hypothesis will be lost? I think not, because there seems to be in mathematics a law of conservation of ideas which (unlike other fields of science) does not allow a good idea to disappear. If

so, then one should not expect a simple and conclusive yes or no answer to the continuum hypothesis. However, an optimist can expect that the future will bring us new insight and knowledge about the continuum (the set of real numbers) which will enrich mathematics without making present knowledge obsolete.

I would like to describe next two impacts on set theory that the independence results of Gödel and Cohen could possibly have. Firstly, I believe that set theory is becoming part of mathematics rather than of metamathematics. By this I also mean that the motives of many of the workers in set theory are the same as those of other mathematicians. They try to solve hard problems, and are interested in their theory for intrinsic reasons. The methods of Gödel and Cohen were developed further into many directions, more questions were asked and new and interesting results were found. Techniques concerned with the constructible universe of Gödel and its variants, forcing methods for obtaining generic extensions, and new axioms became subjects to mathematical investigations. New connections between those directions were found and an active body of research was created. Rather than fulfilling its aim of showing the independence of the continuum hypothesis and they dying at the impasse, the work of Gödel and Cohen proved to be a vital source of inspiration for the development of a coherent area of mathematics. Experience acquired in applying these new methods resulted in theorems and proofs which are not independence results but rather sophisticated results (in ZF alone) concerned with objects whose existence could have been accepted by the first workers in set theory.²

Secondly, I would like to advocate a "liberal" attitude towards the fundamental questions of set theory and mathematics; and I think that the undecidability of the continuum hypothesis supports this attitude. Set theory was born in a storm—one of the most passionate in modern thought. The acceptance of the *infinite* as a legitimate field of study the way Cantor saw it was slow and difficult. The opposition to Cantor's theory and its impact on his life are well known. DAUBEN (1979: p. 162) tells us that "Cantor believed that the unfortunate fate of his own work was the product of a system in which a single individual could ruin any

² I would like to quote the following passage from POINCARÉ (1916: p. 49): "les faits mathématiques dignes d'être étudiés, ce sont ceux qui, par leur analogie avec d'autres faits, sont susceptibles de nous conduire à la connaissance d'une loi mathématique de la même façon que les faits expérimentaux nous conduisent à la connaissance d'une loi physique. Ce sont ceux qui nous révèlent des parentés insoupçonnées entre d'autres faits, connus depuis longtemps, mais qu'on croyait à tort étrangers les uns aux autres."

chances that a young, controversial mathematician might have of gaining recognition . . . The price of “freedom” was isolation and discrimination. According to Cantor, being radical and unorthodox meant that one suffered poverty and recrimination.” Now, it is true that the heated debates on the foundations of mathematics sharpened our intuition on basic questions such as: what is a definition? when does a mathematical object exist? what is an effective process? and so on. However, the personal price paid was too high, and today (being aware to the independence results) we shrug our shoulders when reading the debates concerning the truth or falsity of the axiom of choice for example³.

It is certainly very natural to fight for one’s ideas and convictions and to present one’s point of view as the only possible or the most important. However, I sympathize with Hadamar who wrote [concerning the axiom of choice; see MOORE (1982) appendix 1]: “Consequently, there are two conceptions of mathematics, two mentalities, in evidence. After all that has been said up to this point, I do not see any reason for changing mine. I do not mean to impose it. At the most, I shall note in its favor the [following] arguments’.

I suspect that Cantor himself would not have accepted my point of view. He believed that the infinite cardinal numbers have a reality as convincing as physical objects, and he even speculated on the possibility of applying set-theoretical notions to find a unifying basis to such distinct objects as a Rembrandt portrait and a Beethoven symphony. Perhaps he would have liked to find evidence for his hypothesis in physics (as suggested here yesterday by professor Foreman). Well, I will not object to that; after all didn’t he say that the essence of mathematics is in its freedom?

2. On commutative families of functions: The problem and results

DEFINITIONS. (a) We say that $F = \langle f_i \mid i \in \omega \rangle$ is a commutative family on ω_1 iff $f_i: \omega_1 \rightarrow \omega_1$, and for all i, j the compositions commute: $f_i \circ f_j = f_j \circ f_i$.

(b) Given a commutative family, we say that $X \subset \omega_1$ is *free* iff for all $x \neq y$ in X , and for all $i, f_i(x) \neq y$.

We are concerned here with the question of the existence of uncountable free sets, and we shall show that this question is undecidable: In §3

³ For those debates see FRAENKEL *et al.* (1973) or MOORE (1982) etc.

we give a model in which every commutative family on ω_1 has an uncountable free set. And in a subsequent paper we shall show that in the constructible universe L , there is a commutative family on ω_1 such that every free set is countable. We do not know what is the consequence of the continuum hypothesis on this problem. This work was inspired by J. Steprans' investigation of the number of submodules.

3. Any commutative family has an uncountable free set

An application of SOCA

In ABRAHAM *et al.* (1985), an axiom, SOCA, was introduced and proved consistent with $ZFC + \neg CH$. One consequence of SOCA is the following axiom [proved previously consistent in BAUMGARTNER (1980)].

(*) Any uncountable family of subsets of ω contains an uncountable chain or an uncountable antichain.

Now a *chain* is a family C such that for $a, b \in C$ $a \subseteq b$ or $b \subseteq a$. And an *antichain* is a family A such that $a, b \in A$ and $a \neq b \Rightarrow a \not\subseteq b$ and $b \not\subseteq a$.

3.1. THEOREM. (*) implies that any commutative family of functions on ω_1 has an uncountable free set.

The proof is given in this section. So let $F = \langle f_i \mid i \in \omega \rangle$ be a given commutative family. We assume, for notational convenience that f_0 is the identity function.

DEFINITION. For any $\alpha \in \omega_1$, let $A_\alpha = \{i \in \omega \mid f_i(\alpha) = \alpha\}$.

The following is easy.

LEMMA. Suppose that for some j , $f_j(\alpha) = \beta$. Then for any i $A_{f_j(\alpha)} \subseteq A_{f_j(\beta)}$. And in particular $A_\alpha \subseteq A_\beta$.

DEFINITION. We say that a and b are *obviously free* iff

- (a) $\exists j$ $A_{f_j(\alpha)} \not\subseteq A_{f_j(\beta)}$ and
- (b) $\exists j$ $A_{f_j(\beta)} \not\subseteq A_{f_j(\alpha)}$.

CONCLUSION. If a and b are obviously free then $\{\alpha, \beta\}$ is free.

MAIN LEMMA. *Suppose $X \subset \omega_1$ is uncountable, and does not contain an uncountable free set. Then some $\alpha, \beta \in X$ are obviously free.*

Let us prove Theorem 3.1 assuming first this Lemma. Well, put

$$E_\alpha = \bigcup_{i \in \omega} \{i\} \times A_{f_i(\alpha)}.$$

Then, α and β are obviously free iff $E_\alpha \not\subseteq E_\beta$ and $E_\beta \not\subseteq E_\alpha$. Using a correspondence between $\omega \times \omega$ and ω , we can view E_α as a subset of ω ; and then α and β are obviously free iff $\{E_\alpha, E_\beta\}$ is an antichain.

Now the theorem follows from (*), because the Main Lemma says that $\{E_\alpha \mid \alpha \in \omega_1\}$ cannot contain an uncountable chain, and hence must contain an uncountable antichain which gives a free set.

Now to the proof of the main Lemma.

NOTATION: For $X \subset \omega_1, \tau \in \omega_1, i \in \omega$

$$f_i^{-1}(\tau) \cap X = \{\alpha \in X \mid f_i(\alpha) = \tau\}.$$

The following is easy.

3.2. SUBLEMMA. *Suppose that for some $X \subset \omega_1$ there are uncountably many $\tau \in X$ satisfying*

$$\forall i \mid f_i^{-1}(\tau) \cap X \mid \leq \aleph_0.$$

Then there exists an uncountable free subset of X .

3.3. SUBLEMMA. *If $X \subset \omega_1$ is uncountable, and for all $\alpha, \beta \in X, A_\alpha = A_\beta$, then X contains an uncountable free set.*

PROOF. By Sublemma 3.2 we can find some $\tau \in X$ and i with $F = f_i^{-1}(\tau) \cap X$ uncountable, or else we are done. Now, F must be free. Because otherwise $f_j(\alpha) = \beta$ for some $\alpha \neq \beta$ in F , and so $j \in A_\tau$. Thus $A_\tau \neq A_\alpha$.

Now we prove the Main Lemma. So let $S \subset \omega_1$ be uncountable and with no uncountable free subset. By Sublemma 3.3, $(\forall \alpha \in X)\{\beta \in X \mid A_\alpha = A_\beta\}$ is countable. Hence we can take a subset of X and assume that for any $\alpha \neq \beta$ in $X, A_\alpha \neq A_\beta$. Also there are uncountably many $\tau \in X$ such that, for some $i, f_i^{-1}(\tau) \cap X$ is uncountable (we only need two

elements from this set). For any such $\tau \in X$ choose such an $i < \omega$ and pick distinct $\alpha = \alpha(\tau)$, $\beta = \beta(\tau)$ in $f_i^{-1}(\tau) \cap X$. Since $A_\alpha \neq A_\beta$, for some $k \in \omega$, $k \in A_\alpha - A_\beta$ (say). For uncountably many τ 's the same i was chosen, and the same k was picked in $A_\alpha - A_\beta$. We actually need only two such τ 's. Say τ_0 and τ_1 .

Since $A_{\tau_0} \neq A_{\tau_1}$, we can assume for example that $A_{\tau_1} \not\subseteq A_{\tau_0}$. Put $\alpha_0 = \alpha(\tau_0)$ and $\beta_1 = \beta(\tau_1)$.

Now we claim that α_0 and β_1 are obviously free. Indeed $k \in A_{\alpha_0} - A_{\beta_1}$, so $A_{\alpha_0} \not\subseteq A_{\beta_1}$. And also $\tau_1 = f_i(\beta_1)$, $\tau_0 = f_i(\beta_1)$, and $A_{\tau_1} \not\subseteq A_{\tau_0}$.

References

- ABRAHAM, U., RUBIN, M. and SHELAH, S., 1985, *On the consistency of some partition theorems for continuous colorings, and the structure of \aleph_1 -dense real order types*, *Annals of pure and Applied Logic*, 29, pp. 123–206.
- BAUMGARTNER, J.E., 1980, *Chains and antichains in $P(\omega)$* , *J. Symbolic Logic* 45, pp. 85–92.
- BENACERRAF, P. and PUTNAM, H., ed., 1964, *Philosophy of Mathematics* (Prentice Hall, Englewood Cliffs).
- COHEN, P.J., 1966, *Set Theory and the Continuum Hypothesis* (Benjamin, New York).
- DAUBEN, J.W., 1979, *Georg Cantor* (Harvard University Press, Cambridge).
- FRAENKEL, A., BAR-HILLEL, Y. and LEVY, A., 1973, *Foundation of Set Theory* (North-Holland, Amsterdam).
- GÖDEL, K., 1938, *The consistency of the axiom of choice and of the generalized continuum hypothesis*, *Proc. Natl. Acad. Sci. U.S.A.* 24, pp. 556–557.
- MOORE, G.H., 1982, *Zermelo's Axiom of Choice* (Springer, New York).
- STEPRANS, J., 1984, *The number of submodules*, *London Mathematical Society*, 49 (3), pp. 183–194.

POLARIZED PARTITION RELATIONS AND ALMOST-DISJOINT FUNCTIONS

JAMES E. BAUMGARTNER

*Department of Mathematics and Computer Science, Dartmouth College, Hanover,
NH 03755, U.S.A.*

1. Introduction

Two functions f and g from ω_1 to ω are *almost-disjoint* if $\{\alpha : f(\alpha) = g(\alpha)\}$ is countable. If $\kappa, \lambda, \mu, \nu$ and ρ are cardinals, then the partition relation

$$\binom{\kappa}{\lambda} \rightarrow \binom{\mu}{\nu}_\rho^{1,1}$$

is defined to mean that $\forall f : \kappa \times \lambda \rightarrow \rho \exists A \subseteq \kappa \exists B \subseteq \lambda \mid A \mid = \mu, \mid B \mid = \nu$ and f is constant on $A \times B$. It is not difficult to see that

$$\binom{\kappa}{\omega_1} \not\rightarrow \binom{2}{\omega_1}_\omega^{1,1}$$

is equivalent to the assertion that there is a family F of pairwise almost-disjoint functions from ω_1 to ω with $\mid F \mid = \kappa$. Simply associate $f : \kappa \times \omega_1 \rightarrow \omega$ with $\langle f_\xi : \xi < \kappa \rangle$ where $f_\xi(\alpha) = f(\xi, \alpha)$.

The non-existence of a family F of pairwise almost-disjoint functions from ω_1 to ω with $\mid F \mid = \aleph_2$ is known (DONDER and LEVINSKI 1987) to be of very high consistency strength. It is equivalent to the polarized partition relation

$$\binom{\omega_2}{\omega_1} \rightarrow \binom{2}{\omega_1}_\omega^{1,1}.$$

In Section 2 we begin by showing that

$$\left(\begin{matrix} \omega_2 \\ \omega_1 \end{matrix}\right) \rightarrow \left(\begin{matrix} 2 \\ \omega_1 \end{matrix}\right)_\omega^{1,1} \text{ implies } \left(\begin{matrix} \omega_2 \\ \omega_1 \end{matrix}\right) \rightarrow \left(\begin{matrix} n \\ \omega_1 \end{matrix}\right)_\omega^{1,1} \text{ for all } n < \omega ,$$

a result due independently to Donder and Levinski. In Sections 3 and 4 we show that the occurrence of ω_2 in this result is critical, for under the consistency of ZF alone we obtain (in Section 3) the consistency of ZFC + CH + $2^{\aleph_1} = \aleph_3$ and

$$\left(\begin{matrix} \omega_3 \\ \omega_1 \end{matrix}\right) \rightarrow \left(\begin{matrix} \omega \\ \omega_1 \end{matrix}\right)_\omega^{1,1}$$

as well as (in Section 4) for each $n \geq 2$ the consistency of

$$\left(\begin{matrix} \omega_3 \\ \omega_1 \end{matrix}\right) \rightarrow \left(\begin{matrix} n \\ \omega_1 \end{matrix}\right)_\omega^{1,1} \text{ while } \left(\begin{matrix} \kappa \\ \omega_1 \end{matrix}\right) \not\rightarrow \left(\begin{matrix} n+1 \\ \omega_1 \end{matrix}\right)_\omega^{1,1} .$$

The latter result for $n = 2$ was obtained earlier by Peter Komjáth, who made use of the assumption that the existence of a weakly compact cardinal is consistent.

Throughout the paper we use standard notation. The reader is presumed to be familiar with single-step forcing extensions.

2. An observation for \aleph_2

Let us begin by observing that if we are given a family of \aleph_2 functions from ω_1 to ω that are almost disjoint taken n at a time, then there is such a family that is pairwise almost disjoint. The following result was found independently by Donder and Levinski, and — together with a great deal more information on Chang’s Conjecture-like properties — will appear in DONDER and LEVINSKI (1987)

THEOREM 2.1.

$$\left(\begin{matrix} \omega_2 \\ \omega_1 \end{matrix}\right) \rightarrow \left(\begin{matrix} 2 \\ \omega_1 \end{matrix}\right)_\omega^{1,1} \text{ implies } \left(\begin{matrix} \omega_2 \\ \omega_1 \end{matrix}\right) \rightarrow \left(\begin{matrix} n \\ \omega_1 \end{matrix}\right)_\omega^{1,1} \text{ for all } n < \omega .$$

PROOF. We prove the contrapositive. Suppose $n \geq 3$ and $f: \omega_2 \times \omega_1 \rightarrow \omega$ witnesses $\left(\begin{matrix} \omega_2 \\ \omega_1 \end{matrix}\right) \not\rightarrow \left(\begin{matrix} n \\ \omega_1 \end{matrix}\right)_\omega^{1,1}$. For $\xi < \omega_2$ define $f_\xi: \omega_1 \rightarrow \omega$ by $f_\xi(\alpha) = f(\xi, \alpha)$. Then the f_ξ are almost disjoint taken n at a time, i.e. $\forall x \in [\omega_2]^n \exists \alpha(x) < \omega_1 \forall \beta \geq \alpha$ the values $f_\xi(\beta)$ for $\xi \in x$ are not all the same. We will define

almost-disjoint functions $g_\xi : \omega_1 \rightarrow \omega \times n$ so that $g_\xi(\alpha) = (f_\xi(\alpha), i)$ for some $i < n$. The g_ξ are defined by induction. Let $k_\xi : \omega_1 \rightarrow \xi$ be onto, and let $C_\xi = \{\alpha < \omega_1 : \forall x \in [k_\xi \text{ ``}\alpha\text{''}]^{n-1} \alpha(x \cup \{\xi\}) < \alpha\}$. Then C_ξ is closed and unbounded in ω_1 . To define $g_\xi(\alpha)$, let $\delta = \max(C \cap \alpha)$ and note that there are at most $n - 1$ elements η of $k_\xi \text{ ``}\delta\text{''}$ such that $f_\eta(\alpha) = f_\xi(\alpha)$. Since each $g_\eta(\alpha)$ has the form $(f_\eta(\alpha), j)$ for some $j < n$ we may choose $i < n$ distinct from all such j , and we set $g_\xi(\alpha) = (f_\xi(\alpha), i)$. Now if $\eta < \xi$ and $\delta \in C_\xi$ is large enough so that $\delta > \eta$, then $\forall \alpha \geq \delta$ we must have $g_\xi(\alpha) \neq g_\eta(\alpha)$. But the g_ξ easily give rise to a witness g to

$$\binom{\omega_2}{\omega_1} \not\rightarrow \binom{2}{\omega_1}_\omega^{1,1}.$$

3. A dry run

The main result of this paper is the fact that it is consistent with $\text{CH} + 2^{\aleph_1} \geq \aleph_3$ that Theorem 2.1 fails badly when ω_2 is replaced by ω_3 . First, however, to illustrate the forcing method we present a somewhat easier proof that may be of interest in its own right, namely it is consistent with $\text{CH} + 2^{\aleph_1} \geq \aleph_3$ that there is no family of \aleph_3 pairwise almost-disjoint functions from ω_1 to ω .

The following combinatorial lemma will be of use in both results.

LEMMA 3.1. *Suppose α is an ordinal, $\alpha \geq 2$, and ${}^\omega\alpha = \bigcup \{A_m : m \in \omega\}$. Then $\exists k \exists \sigma \in {}^k\alpha \exists m \forall \beta < \alpha \exists f_\beta \in A_m \sigma \subseteq f_\beta$ and $f_\beta(k) = \beta$.*

(One can think of the f_β as forming a complete split at σ .)

PROOF. Suppose not. Then we may construct a sequence $\sigma_n \in {}^n\alpha$ such that $\sigma_n \subseteq \sigma_{n+1}$ and $\forall f \in A_n$ it is not the case that $\sigma_{n+1} \subseteq f$. But if $f = \bigcup \{\sigma_n : n \in \omega\}$ and $f \in A_m$, then $\sigma_{m+1} \subseteq f$, a contradiction.

The next theorem implies the consistency with 2^{\aleph_1} large of the assertion that there is no family of \aleph_3 pairwise almost-disjoint functions from ω_1 to ω .

THEOREM 3.2. *Suppose $\kappa \geq \aleph_3$ and cf $\kappa > \omega_1$. Then if ZF is consistent so is*

$$\text{ZFC} + \text{CH} + 2^{\aleph_1} = \kappa + \binom{\omega_3}{\omega_1} \rightarrow \binom{\omega}{\omega_1}_\omega^{1,1}.$$

PROOF. We will begin with a model of GCH and add κ Cohen subsets of ω_1 . The partial ordering P of forcing conditions consists of all countable partial functions from κ to 2. It is well known that P is countably closed and (by CH) has the \aleph_2 -chain condition, and that $2^{\aleph_1} = \kappa$ in the extension. It will suffice to check the polarized partition relation in the extension.

We must show that if $\langle f_\xi : \xi < \omega_3 \rangle$ are functions from ω_1 to ω then countably many of them must agree uncountably often. Let \dot{f}_ξ be a forcing term for f_ξ . Standard arguments show that for each ξ there is $Z(\xi) \subseteq \kappa$ of cardinality \aleph_1 such that $\forall p \in P \forall \alpha < \omega_1 \forall n$ if $p \Vdash \dot{f}_\xi(\alpha) = n$ then $p \upharpoonright Z(\xi) \Vdash \dot{f}_\xi(\alpha) = n$. The usual way of constructing $Z(\xi)$ is to choose maximal antichains A_α of conditions p forcing $\dot{f}_\xi(\alpha) = n$ for some n , and to let $Z(\xi) = \bigcup \{ \bigcup \{ \text{dom } p : p \in A_\alpha \} : \alpha < \omega_1 \}$. Let $P_\xi = \{ p \in P : \text{dom } p \subseteq Z(\xi) \}$.

By $2^{\aleph_1} = \aleph_2$ we may find $X \subseteq \omega_3$ of cardinality \aleph_3 such that $\langle Z(\xi) : \xi \in X \rangle$ is a Δ -system with kernel Δ and whenever $\xi, \eta \in X$ there is a (unique) order-preserving mapping $\pi_{\xi\eta}$ from $Z(\xi)$ to $Z(\eta)$ that is the identity on Δ . We may also assume that $\pi_{\xi\eta}$ lifts to an isomorphism of P_ξ with P_η in the obvious way, and that $\forall p \in P_\xi \forall \alpha < \omega_1 \forall n < \omega p \Vdash \dot{f}_\xi(\alpha) = n$ iff $\pi_{\xi\eta}(p) \Vdash \dot{f}_\eta(\alpha) = n$.

Fix $\zeta \in X$. A sequence $\langle p_n : n < \omega \rangle$ of elements of P_ζ is said to be *consistent mod Δ* if $p_n \upharpoonright \Delta$ is the same for all $n < \omega$. Note that if $\langle p_n : n < \omega \rangle$ is consistent mod Δ and $\langle \xi_n : n < \omega \rangle$ is a sequence of distinct elements of X , then $\bigcup \{ \pi_{\zeta\xi_n}(p_n) : n < \omega \}$ is a condition, i.e., it lies in P .

LEMMA 3.3 (Main Lemma). *There is $p \in P_\zeta$ such that $\forall \beta < \omega_1 \forall \langle p_n : n < \omega \rangle$ if $\langle p_n : n < \omega \rangle$ is consistent mod Δ and $p_n \leq p$ for all n , then $\exists \alpha > \beta \exists m < \omega \exists \langle q_n : n < \omega \rangle \langle q_n : n < \omega \rangle$ is consistent mod Δ and for all n , $q_n \leq p_n$ and $q_n \Vdash \dot{f}_\zeta(\alpha) = m$.*

PROOF. Suppose not. We will build a tree of conditions p_σ in P_ζ (and ordinals $\beta_\sigma < \omega_1$) indexed by $\sigma \in \bigcup \{ {}^n\omega : n \in \omega \}$ such that for each σ, β_σ and $\langle p_{\sigma i} : i < \omega \rangle$ form a counterexample showing that p_σ does not satisfy the lemma. (Here σi denotes the concatenation of σ with $\langle i \rangle$.) We will also have $p_\sigma \upharpoonright \Delta = p_\tau \upharpoonright \Delta$ whenever $|\sigma| = |\tau|$.

Let $p_{\langle \rangle}$ be arbitrary. Suppose p_σ has been found for all $\sigma \in {}^n\omega$. Let $\langle \sigma_k : k < \omega \rangle$ enumerate ${}^n\omega$. First we define β_{σ_k} and a sequence $\langle q_{\sigma_k i} : i < \omega \rangle$ by induction on k as follows. For $\sigma = \sigma_0$ let β_σ and $\langle q_{\sigma i} : i < \omega \rangle$ be any counterexample witnessing the failure of the lemma for p_σ . If $\sigma = \sigma_{k+1}$ then let β_σ and $\langle q_{\sigma i} : i < \omega \rangle$ witness the failure of the lemma for

$p_\sigma \cup (q_{\sigma k} \mid \Delta)$ where i is any element of ω . Thus if $k < l$ and $i, j \in \omega$ are arbitrary we always have $q_{\sigma j} \mid \Delta \leq q_{\sigma k} \mid \Delta$. Finally, for each $\sigma \in {}^n\omega$ and $i \in \omega$ let $p_{\sigma i} = q_{\sigma i} \cup (\cup \{q_{\sigma k} \mid \Delta : k < \omega\})$. This completes the construction of the tree. Note that for all σ, β_σ and $\langle p_{\sigma i} : i < \omega \rangle$ still witness the failure of the lemma for p_σ .

Let $\langle g_\alpha : \alpha < \omega_1 \rangle$ enumerate ${}^\omega\omega$. For each α , let $p_\alpha = \cup \{p_{g_\alpha \mid n} : n \in \omega\}$. Note that $p_{\alpha_1} \mid \Delta = p_{\alpha_2} \mid \Delta$ whenever $\alpha_1, \alpha_2 < \omega_1$. Choose $\beta < \omega_1$ so large that $\beta > \beta_\sigma$ for all $\sigma \in \cup \{{}^n\omega : n < \omega\}$. By induction on α we determine $q_\alpha \leq p_\alpha$ so that $q_\alpha \leq \cup \{q_\gamma \mid \Delta : \gamma < \beta\}$ and for some $m_\alpha, q_\alpha \Vdash \dot{f}_\zeta(\beta) = m_\alpha$. Let $A_m = \{g_\alpha : m_\alpha = m\}$. By Lemma 3.1 there exist $k, \sigma \in {}^k\omega, m < \omega$ and $\langle \alpha_i : i < \omega \rangle$ such that $\alpha_i \in A_m$ for all i and $q_{\alpha_i} \leq p_{\sigma i}$. But now if $r_i = q_{\alpha_i} \cup \cup \{q_{\alpha_j} \mid \Delta : j \in \omega\}$ then $r_i \leq p_{\sigma i}, r_i \Vdash \dot{f}_\zeta(\beta) = m, \langle r_i : i \in \omega \rangle$ is consistent mod Δ , and $\beta > \beta_\sigma$, all of which contradicts the fact that $\beta_\sigma, \langle p_{\sigma i} : i < \omega \rangle$ witness the failure of the lemma for p_σ . This contradiction establishes the Main Lemma.

Now we can finish the proof of Theorem 3.2. Let $p \in P_\zeta$ be as in the Main Lemma, and let $\xi_i, i < \omega$, be distinct elements of X . Let $\bar{p} = \cup \{\pi_{\zeta \xi_i}(p) : i < \omega\}$. We claim \bar{p} forces the f_{ξ_i} to agree uncountably often. Let $q \leq \bar{p}$ and $\beta < \omega_1$ be arbitrary. Let $p_i = \pi_{\zeta \xi_i}(q \mid Z(\xi_i))$. Then $\langle p_i : i < \omega \rangle$ is consistent mod Δ and $p_i \leq p$ for all i . Hence there are $\alpha > \beta, m < \omega$ and $\langle q_i : i < \omega \rangle$ as in the Main Lemma. But then $r = \cup \{\pi_{\zeta \xi_i}(q_i) : i < \omega\}$ has the property that $r \Vdash \dot{f}_{\xi_i}(\alpha) = m$ for all $i < \omega$, since $q_i \Vdash \dot{f}_{\xi_i}(\alpha) = m$ and $r \leq \pi_{\zeta \xi_i}(q_i) \Vdash \dot{f}_{\xi_i}(\alpha) = m$. Of course $r \in P$ since $\langle q_i : i < \omega \rangle$ is consistent mod Δ .

Note that we could have modified the Main Lemma to prove that p could be found below any given condition (which is what is really needed) but that seems to introduce unnecessary complication.

4. The main result

Beginning with a model of GCH with κ as in Theorem 3.2, it is completely straightforward to define a forcing notion that produces a generic function $f : \kappa \times \omega_1 \rightarrow \omega$ witnessing

$$\left(\begin{matrix} \kappa \\ \omega_1 \end{matrix} \right) \not\rightarrow \left(\begin{matrix} 2 \\ \omega_1 \end{matrix} \right)_\omega^{1,1}.$$

Details are left to the reader. This result is actually the case $n = 1$ in the

following theorem, which combines both approaches. Theorem 4.1 was first proved for the case $n = 2$ in unpublished work by Peter Komjáth, assuming the consistency of the existence of a weakly compact cardinal. Note that Theorem 4.1 shows that in Theorem 2.1 the use of ω_2 is essential.

THEOREM 4.1. *Suppose $\kappa \geq \aleph_3$, cf $\kappa > \omega_1$ and $1 \leq n < \omega$. Then if ZF is consistent so is*

$$ZFC + CH + 2^{\aleph_1} = \kappa + \binom{\omega_3}{\omega_1} \rightarrow \binom{n}{\omega_1}_{\omega}^{1,1} + \binom{\kappa}{\omega_1} \not\rightarrow \binom{n+1}{\omega_1}_{\omega}^{1,1}.$$

PROOF. Assume $n \geq 2$. The proof is very similar in structure to the proof of Theorem 3.2 but it is a little more complicated, partly because the “obvious” partial ordering to use does not work.

Let P be the set of all functions of the form $p: A \times \alpha \rightarrow \omega$, where A is some countable subset of κ and $\alpha < \omega_1$. We write $A = A(p)$, $\alpha = \alpha(p)$. If we wish P to adjoin a generic witness to

$$\binom{\kappa}{\omega_1} \not\rightarrow \binom{n+1}{\omega_1}_{\omega}^{1,1},$$

then the most natural requirement to make in defining the ordering on P is to say that $p \leq q$ iff

$$\forall x \in [A(q)]^{n+1} \forall \beta < \alpha(p) \text{ if } \beta \geq \alpha(q) \text{ then } p \text{ is not constant on } x \times \{\beta\}. \tag{P1}$$

Let us verify that (P1) is not enough. Suppose $f: \kappa \times \omega_1 \rightarrow \omega$ is adjoined by P using only condition (P1). If A is any countable subset of κ then it is clear by genericity that

$$X = \{ \alpha : \forall m < \omega \{ \xi \in A : f(\xi, \alpha) = m \} = n - 1 \}$$

is uncountable. Suppose $\xi, \eta \in \kappa - A$, $\xi \neq \eta$. Then there is $\beta < \omega_1$ such that $\forall \alpha > \beta \forall x \in [A \cup \{ \xi, \eta \}]^{n+1}$ it is not the case that f is constant on $x \times \{ \alpha \}$. But if $\alpha > \beta$ and $\alpha \in X$ then we cannot have $f(\xi, \alpha) = f(\eta, \alpha) = m$, say, since $\{ \zeta \in A : f(\zeta, \alpha) = m \}$ already has cardinality $n - 1$. Hence if f_ξ and f_η are defined from f as usual it is clear that $f_\xi \upharpoonright X$ and $f_\eta \upharpoonright X$ are

almost disjoint, and this gives rise to a function witnessing

$$\binom{\kappa}{\omega_1} \not\rightarrow \binom{2}{\omega_1}_\omega^{1,1}.$$

One solution is to make the following additional requirement. Let U be a non-principal ultrafilter on ω . Then for $p, q \in P$ we say $p \leq q$ iff (P1) holds and also

$$\forall \beta < \alpha(p) \text{ if } \beta \geq \alpha(q) \text{ then } p''(A(q) \times \{\beta\}) \notin U. \quad (\text{P2})$$

Assume GCH as above. It is easy to see that P is countably closed and has the \aleph_2 -chain condition, and that

$$\binom{\kappa}{\omega_1} \not\rightarrow \binom{n+1}{\omega_1}_\omega^{1,1}$$

holds in the extension.

Suppose $f: \omega_3 \times \omega_1 \rightarrow \omega$ in the extension. Let $f_\xi(\alpha) = f(\xi, \alpha)$ as before. We must find n of the f_ξ that agree uncountably often. Let \dot{f}_ξ be a name for f_ξ . Then as before we can find $Z(\xi) \subseteq \kappa, |Z(\xi)| = \aleph_1$, and $\forall p \in P \forall \alpha < \omega_1 \forall m < \omega$ if $p \Vdash \dot{f}_\xi(\alpha) = m$ then $p \upharpoonright (Z(\xi) \times \omega_1) \Vdash \dot{f}_\xi(\alpha) = m$. As before we may find a set $X \subseteq \omega_3$ of cardinality \aleph_3 so that $\langle Z(\xi): \xi \in X \rangle$ is a Δ -system with kernel Δ and for all $\xi, \eta \in X$ there is a (unique) order-preserving bijection from $Z(\xi)$ to $Z(\eta)$ that lifts canonically to an isomorphism between P_ξ and P_η , where $P_\xi = \{p \in P: \text{dom } p \subseteq Z(\xi) \times \omega_1\}$.

It remains now only to state, prove and apply the analogue of the Main Lemma.

Fix $\zeta \in X$. A sequence $\{p_i: i < n\}$ of elements of P_ζ is *consistent mod* Δ if for all $i, j < n$ we have $p_i \upharpoonright (\Delta \times \omega_1) = p_j \upharpoonright (\Delta \times \omega_1)$, $A(p_i) = A(p_j)$ and $\alpha(p_i) = \alpha(p_j)$. We say $\langle p_i: i < n \rangle$ is *consistent mod* (Δ, A, α) if it is consistent mod Δ , $A \subseteq A(p_i)$ and $\alpha \leq \alpha(p_i)$ for $i < n$, and in addition

$$\begin{aligned} & \forall \beta \geq \alpha \forall m < \omega \\ & |\{\gamma \in \Delta \cap A: p_i(\gamma, \beta) = m\}| \\ & + \sum_{i < n} |\{\gamma \in A - \Delta: p_i(\gamma, \beta) = m\}| \leq n. \end{aligned}$$

We say $\langle p_i : i < n \rangle \leq \langle q_i : i < n \rangle$ if $p_i \leq q_i$ for all $i < n$, $\langle q_i : i < n \rangle$ is consistent mod Δ and $\langle p_i : i < n \rangle$ is consistent mod $(\Delta, A(q_i), \alpha(q_i))$ where $i < n$. This notion is intended to guarantee that an amalgamation of copies of the p_i 's (in distinct P_{ξ_i} 's) extends the corresponding amalgamation of copies of the q_i 's. Note that \leq for n -tuples is a partial ordering. Let $c(p)$ be the constant n -tuple $\langle p, p, \dots, p \rangle$.

LEMMA 4.2 (Main Lemma). *There is $p \in P_\zeta$ such that $\forall \beta < \omega_1 \forall \langle p_i : i < n \rangle$ if $\langle p_i : i < n \rangle \leq c(p)$ then $\exists \alpha > \beta \exists m < \omega \exists \langle q_i : i < n \rangle \leq \langle p_i : i < n \rangle \forall i q_i \Vdash \dot{f}_\zeta(\alpha) = m$.*

PROOF. Suppose not. Once again we build a tree $\langle p_\sigma : \sigma \in \bigcup \{^k n : k \in \omega\} \rangle$ of elements of P_ζ together with ordinals $\beta_\sigma < \omega_1$ such that for all σ, β_σ and $\langle p_{\sigma_i} : i < n \rangle$ witness that p_σ does not satisfy the lemma. We always have $p_\tau \leq p_\sigma$ when $\sigma \subseteq \tau$ and we will have A_k, α_k so that whenever $\sigma \in {}^k n$ then $A(p_\sigma) = A_k, \alpha(p_\sigma) = \alpha_k$. Furthermore we will insist that for any sequence $\langle \tau_i : i < n \rangle$ of distinct elements of ${}^{k+1}n$, $\langle p_{\tau_i} : i < n \rangle$ is consistent mod (Δ, A_k, α_k) .

Let $p_{\langle \rangle}$ be arbitrary. Suppose we have constructed p_σ for all $\sigma \in {}^k n$. Let $\langle \sigma_j : j < l \rangle$ enumerate ${}^k n$. By induction on $m < l$ we will show how to find $\langle p_{\sigma_j} : i < n \rangle$ that satisfy our requirements for all $j \leq m$. For $m = 0$ this is easy, using the assumption that the lemma is false. Suppose we have found $\langle p'_{\sigma_j} : i < n \rangle$ which work for all $j \leq m$. Let $\sigma = \sigma_{m+1}$.

LEMMA 4.3. *There is $q \in P_\zeta$ such that $c(q) \leq c(p_\sigma)$ and every n -tuple of distinct elements of $\{p'_{\sigma_j} : j \leq m, i < n\} \cup \{q\}$ is consistent mod (Δ, A_k, α_k) .*

PROOF. Note that saying $c(q) \leq c(p_\sigma)$ is stronger than saying $q \leq p_\sigma$. Let A, α be the common values of $A(p'_{\sigma_j}), \alpha(p'_{\sigma_j})$. We must define q on $A \times \alpha$. Suppose $(\xi, \beta) \in A \times \alpha$. If $(\xi, \beta) \in A_k \times \alpha_k$ then let $q(\xi, \beta) = p_\sigma(\xi, \beta)$. If $\xi \in \Delta \cap A$ let $q(\xi, \beta) = p'_{\sigma_j}(\xi, \beta)$. If $\xi \in A - \Delta$ and $\beta < \alpha_k$ then $q(\xi, \beta)$ may be chosen arbitrarily. That leaves the case $\xi \in A - \Delta, \beta \geq \alpha_k$. If $\xi \notin A_k$ then $q(\xi, \beta)$ may be chosen arbitrarily, so suppose $\xi \in A_k - \Delta$. In this case we know, since $p'_{\sigma_j} \leq p_{\sigma_j}$, that if Z_{j_i} is the range of p'_{σ_j} on $A_k \times \{\beta\}$ then by (P2) $Z_{j_i} \not\subseteq U$. Hence we may find infinite $Y \subseteq \omega, Y \not\subseteq U, Y$ disjoint from all the Z_{j_i} . Define q on $(A_k - \Delta) \times \{\beta\}$ so that it is one-to-one with range Y . Then $q \leq p_\sigma$ and $c(q)$ is consistent

$\text{mod}(\Delta, A_k, \alpha_k)$, so $c(q) \leq c(p_\sigma)$. It is now easy to check that every n -tuple of distinct elements of $\{p'_{\sigma i} : j \leq m, i < n\} \cup \{q\}$ is consistent $\text{mod}(\Delta, A_k, \alpha_k)$.

Next find $\langle p_{\sigma i} : i < n \rangle$ and β_σ so that $\langle p_{\sigma i} : i < n \rangle \leq c(q)$ and $\beta_\sigma, \langle p_{\sigma i} : i < n \rangle$ together witness that the Main Lemma fails for q , hence for p_σ . Of course the $p_{\sigma i}$ now have a larger domain, say $B \times \gamma$, then the $p'_{\sigma i}$, but by a process like the proof of Lemma 4.3 it is easy to enlarge the domain of each $p'_{\sigma i}$ to $B \times \gamma$, and the resulting condition is $p_{\sigma i}$. Thus we have been able to construct $\langle p_{\sigma i} : i < n \rangle$ for $j \leq m + 1$ and the induction is complete.

This also completes the inductive construction of the tree p_σ of elements of P_ζ .

Let $\langle g_\alpha : \alpha < \omega_1 \rangle$ enumerate ${}^\omega n$. For each α let $p_\alpha = \bigcup \{p_{g_\alpha | m} : m < \omega\}$. Note that $p_\alpha \upharpoonright (\Delta \times \omega_1)$ is independent of α . Let $A_\omega = \bigcup \{A_n : n \in \omega\}$ and let $\alpha_\omega = \sup\{\alpha_n : n \in \omega\}$. Let $\beta < \omega_1$ be so large that $\beta > \beta_\sigma$ for all $\sigma \in \bigcup \{{}^k n : k \in \omega\}$. Now we construct sequences $\langle p'_\alpha : \alpha < \omega_1 \rangle$ and $\langle q_\alpha : \alpha < \omega_1 \rangle$ by induction so that $p'_\alpha \leq p_\alpha \cup \bigcup \{q_\beta : \beta < \alpha\}$ and for some $m_\alpha, p'_\alpha \upharpoonright \dot{f}_i(\beta) = m_\alpha$. Given p'_γ we must define q_γ . The purpose of choosing $p'_\gamma \leq q_\gamma$ is to ensure that distinct $p'_{\gamma_i}, i < n$, can be extended to r_i so that $\langle r_i : i < n \rangle$ is consistent $\text{mod}(\Delta, A_\omega, \alpha_\omega)$.

Let $A = A(p'_\gamma)$ and let $\alpha = \alpha(p'_\gamma)$. We define q_γ so that its domain is $A \times \alpha$ and $\bigcup \{q_\beta : \beta < \gamma\} \subseteq q_\gamma$. Let q_γ agree with p'_γ on $A \times \sup\{\alpha(q_\nu) : \nu < \gamma\}$ and on $(A \cap \Delta) \times \alpha$. Let δ be such that $\sup\{\alpha(q_\nu) : \nu < \gamma\} \leq \delta < \alpha$. If Z_δ is the range of p'_γ on $A_\omega \times \{\delta\}$ then $Z_\delta \not\subseteq U$ (since $p'_\gamma \leq p_\gamma$) so we can find $Y_\delta \not\subseteq U$, infinite and disjoint from Z_δ . Let q_γ be one-to-one from $(A_\omega - \Delta) \times \{\delta\}$ to Y_δ , and let q_γ be arbitrary on $(A - (A_\omega \cup \Delta)) \times \{\delta\}$.

This completes the construction of the p'_α and q_α . Note that if $\alpha_0 < \dots < \alpha_{n-1}$ and $\alpha_\omega \leq \delta < \alpha(p'_{\alpha_0})$, then for $0 < i < n$ the functions p'_{α_i} all agree with q_{α_0} on $A_\omega \times \{\delta\}$, and either they agree with p'_{α_0} as well or else (in case δ first occurred in the domain of p'_{α_0}) the ranges of p'_{α_0} and q_{α_0} are disjoint on $(A_\omega - \Delta) \times \{\delta\}$. In either case it is straightforward to see that the p'_{α_i} may be extended to r_i such that $\langle r_i : i < n \rangle$ is consistent $\text{mod}(\Delta, A_\omega, \alpha_\omega)$.

By Lemma 3.1 there are $\alpha_0 < \dots < \alpha_{n-1} < \omega_1$ so that all the $m_{\alpha_i} = m$, say, and for some σ we have $p_{\sigma i} \leq p'_{\alpha_i}$ for all i . But if $\langle r_i : i < n \rangle$ is as above then we claim $\langle p_{\sigma i} : i < n \rangle \leq \langle r_i : i < n \rangle$, i.e., that $\langle r_i : i < n \rangle$ is consistent $\text{mod}(\Delta, A_{k+1}, \alpha_{k+1})$, where $\sigma \in {}^k n$. But this follows quickly from the fact that for all $l < k$, if $\tau_i \in {}^{l+1} n$ and $r_i \leq p_{\tau_i}$ then the τ_i are distinct and by our construction $\langle p_{\tau_i} : i < n \rangle$ is consistent $\text{mod}(\Delta, A_l, \alpha_l)$.

Of course $\langle r_i : i < n \rangle$ shows that $\{p_{\sigma_i} : i < n\}$ was not chosen correctly, and this establishes the Main Lemma.

To complete the proof of Theorem 4.1, we proceed exactly as in Theorem 3.2. Let $p \in P_\zeta$ be as in the Main Lemma, and let $\xi_i, i < n$, be distinct elements of X . Let $\bar{p} = \bigcup \{\pi_{\xi_i}(p) : i < n\}$. Let $q \leq \bar{p}$ and $\beta < \omega_1$ be arbitrary, and let $p_i = \pi_{\xi_i}(q \upharpoonright (Z(\xi_i) \times \omega_1))$. By extending q , if necessary, we may assume that $\langle p_i : i < n \rangle$ is consistent mod Δ , and since $q \leq \bar{p}$ we have $\langle p_i : i < n \rangle \leq c(p)$. Hence there are $\alpha > \beta, m < \omega$ and $\langle q_i : i < n \rangle$ as in the Main Lemma. But then $r = \bigcup \{\pi_{\xi_i}(q_i) : i < n\}$ has the property that $r \Vdash \dot{f}_{\xi_i}(\alpha) = m$ for all $i < n$, and we are done.

It is clear that these methods can be generalized to treat regular cardinals larger than ω_1 . Furthermore, for a fixed cardinal like ω_1 it appears to be possible to get versions of Theorem 4.1 for several different values of n holding simultaneously. For example, one might have

$$\binom{\omega_3}{\omega_1} \rightarrow \binom{2}{\omega_1}^{1,1}, \quad \binom{\omega_4}{\omega_1} \rightarrow \binom{3}{\omega_1}^{1,1}, \quad \binom{\omega_5}{\omega_1} \rightarrow \binom{4}{\omega_1}^{1,1},$$

while $2^{\aleph_1} = \aleph_{10}$ and all these relations are best possible. One can also tinker with the value of 2^{\aleph_0} by adjoining Cohen reals after the smoke has cleared in the above result. All these elaborations are left to the reader.

Reference

DONDER, H.-D. and LEVINSKI, J.-P., 1987, *Some principles related to the Conjecture of Chang*, handwritten manuscript.

A DILWORTH DECOMPOSITION THEOREM FOR λ -SUSLIN QUASI-ORDERINGS OF \mathbb{R}

MATTHEW FOREMAN

Ohio State University, Columbus, Ohio 43210, USA

0. Introduction

In HARRINGTON *et al.* (unpublished), it is shown that if \leq is a Borel quasi-ordering of \mathbb{R} with no perfect set of incomparable elements then

- (1) $\mathbb{R} = \bigcup_{n \in \omega} X_n$ where each X_n is \leq -linearly ordered and Borel
- (2) There is a strictly ordered preserving map $F: (\mathbb{R}, \leq) \rightarrow (2^\alpha, <_{\text{lex}})$ for some $\alpha < \omega_1$. ($<_{\text{lex}}$ is the lexico-graphical ordering on functions from α into 2).

This paper contains the proofs of the following facts:

Assume $AD + ZF + DC$. Let \leq be a Suslin, co-Suslin quasi-ordering of \mathbb{R} not having a perfect set of incomparable elements. Suppose λ_1, λ_2 are minimal such that \leq is λ_1 -Suslin and $\not\leq$ is λ_2 -Suslin. Let $T_{\leq} \subseteq (\omega \times \omega \times \lambda_1)^{<\omega}$ and $T_{\not\leq} \subseteq (\omega \times \omega \times \lambda_2)^{<\omega}$ witness this. Let $\lambda = \max(\lambda_1, \lambda_2)$.

Then

THEOREM 1.1. *There are $\langle X_\alpha : \alpha \in \lambda \rangle$ such that:*

- (1) X_α is Suslin
- (2) X_α is pre-linearly ordered by \leq .

THEOREM 1.2. *Let $\kappa > \lambda$ be the least ordinal such that $L_\kappa(\mathbb{R} \cup \{T_{\leq}, T_{\not\leq}\})$ is admissible. Then there is an $\alpha < \kappa$ and an order preserving function $F: \langle \mathbb{R}, \leq \rangle \rightarrow (\{f \mid f: \alpha \rightarrow 2\}, \leq_{\text{lex}})$ such that if $x \not\leq y$ then $F(x) \not\leq F(y)$.*

We deduce the following corollaries in *ZFC*:

Suppose there is a supercompact cardinal and $NS_{\omega_2}[\text{cof}(\omega_1)]$ is ω_3 -saturated and $\leq \in L(\mathbb{R})$ then:

COROLLARY 1. *If \leq is a prelinear ordering of \mathbb{R} then there is no \leq -increasing sequence of reals of length ω_2 .*

COROLLARY 2. *If \leq is a quasi ordering of \mathbb{R} with no perfect set of incomparable elements then $\mathbb{R} = \bigcup_{\alpha \in \omega_1} X_\alpha$ where each X_α is pre-linearly ordered by \leq .*

To deduce these corollaries we use a theorem of FOREMAN and MAGIDOR (in preparation).

THEOREM (ZFC). *If there is a supercompact cardinal and $NS_{\omega_2}[\text{cof}(\omega_1)]$ is ω_3 -saturated then $\theta^{L(\mathbb{R})} < \omega_2$.*

$$(\theta^{L(\mathbb{R})} = \sup\{\alpha : \exists f \in L(\mathbb{R}) \quad f: \mathbb{R} \xrightarrow{\text{onto}} \alpha\})$$

By results of Martin-Steel-Woodin, if there is a supercompact cardinal then there is an inner-model $M \supseteq L(\mathbb{R})$, $M \models ZF + AD. + DC. +$ every set of reals in $L(\mathbb{R})$ is θ -Suslin. The corollaries follow immediately. (In fact, Woodin can construct an inner model $M \models "ZF + AD_{\mathbb{R}} +$ every set of reals is λ -Suslin for some $\lambda < \theta^M"$ from a supercompact. The results of FOREMAN and MAGIDOR (in preparation) show that $\theta^M < \omega_2$, hence, the corollaries hold for M as well.)

The proofs of theorems 1.1 and 1.2 can be viewed as merging the proofs of HARRINGTON *et al.* (unpublished) with the techniques of HARRINGTON and SAMI (1978).

There are some difficulties, as we replace forcing by games. This necessitates §4 and some wrinkles in §5 and §6.

Also used are results of MOSCHOVAKIS (1980) and MARTIN (1983) on the extent of scales and the results of HARRINGTON and KECHRIS (1972) on the determinacy of ordinal games.

It is clear that this theorem is subject to analysis of what amount of determinacy is necessary to prove this theorem for \leq in particular pointclasses. This is omitted as this paper is mostly meant to tie in with FOREMAN and MAGIDOR (in preparation).

We follow the notations of HARRINGTON *et al.*, (unpublished) and HARRINGTON and SAMI (1978). If $s_0 \neq s_1 \in X^{\leq \omega}$ we let $s_0 \wedge s_1 = s_0 \upharpoonright n$ where $n \in \omega$ is largest so that $s_0 \upharpoonright n = s_1 \upharpoonright n$. If \leq is a quasi-ordering, we use \geq for its converse. We use $x < y$ if $x \leq y$ but $y \not\leq x$. For $A \subseteq \mathbb{R} \times \mathbb{R}$ we let $\pi_0(A)$, $\pi_1(A)$ be the projections of A onto the first and second coordinates respectively.

The author would like to thank Donald Martin, Menachem Magidor and David Marker for very useful conversation and correspondence related to this paper.

We call $\leq \subseteq \mathbb{R} \times \mathbb{R}$ a *quasi-ordering* iff it is a reflexive, transitive binary relation. It is *thin* iff there is no perfect set $P \subseteq \mathbb{R}$ such that if $x \neq y$, $x, y \in P$ then $x \not\leq y$.

We now enumerate some descriptive set-theoretic facts we will use without proof.

For the sequel we assume $AD + DC$. Let \leq be λ_1 -Suslin with λ_1 minimal and $\not\leq$ be λ_2 -Suslin with λ_2 minimal and fix trees $T_{\leq} \subseteq (\omega \times \omega \times \lambda_1)^{<\omega}$ and $T_{\not\leq} \subseteq (\omega \times \omega \times \lambda_2)^{<\omega}$ witnessing this. Let $\lambda = \max\{\lambda_1, \lambda_2\}$. Let $\kappa > \lambda_1, \lambda_2$ be minimal such that $N = L_\kappa(\mathbb{R} \cup \{T_{\leq}, T_{\not\leq}\})$ is admissible.

Let Γ be the collection of all $X \subseteq (\lambda \cup \mathbb{R})^n$ ($n \in \omega$) that are Σ_1 definable over N with parameters in $\lambda \cup \{\lambda, \mathbb{R}, T_{\leq}, T_{\not\leq}\}$. Note that Γ is a ‘‘light-face’’ class. Let $\Delta = \Gamma \cap \check{\Gamma}$.

The original proof of propositions 1 and 2 used $AD_{\mathbb{R}}$.

The author is indebted to Donald Martin for remarking that the amount of determinacy necessary for the games in this paper is known to follow from A.D. Specifically, let Γ_0 be the least adequate pointclass of \mathbb{R} containing the scales corresponding to $T_{\leq}, T_{\not\leq}$. By the *Second periodicity Theorem* (MOSCHOVAKIS 1980, p. 311), the least pointless Γ_ω containing Γ_0 and closed under universal and existential real quantification has the property that for all $P \in \Gamma_\omega$ there is a scale on P in Γ_ω .

In particular, since λ is chosen to be minimal there is a norm $\varphi: \mathbb{R} \xrightarrow{\text{onto}} \lambda$ in Γ_ω with a scale on φ in Γ_ω .

Let Γ' be the class of subsets of \mathbb{R} that are absolutely positive inductive over Γ_ω (see MOSCHOVAKIS 1980, p. 410). Then by a theorem of MOSCHOVAKIS (1980, p. 411) Γ' has scales.

Hence there is a pointclass with the scale property that contains a norm $\varphi: \mathbb{R} \xrightarrow{\text{onto}} \lambda$. Consequently the theorems of HARRINGTON-KECHRIS (1972) apply to show that if $A \subseteq \lambda^\omega$ and G_A is the game where the players play ordinals $< \lambda$ then G_A is determined.

Specifically, let $\phi: \mathbb{R} \xrightarrow{\text{onto}} \lambda \cup \{\infty\}$ be a Γ norm. (Such a norm exists by the minimality of λ .) Let G_A be the game:

$$\begin{array}{rcc} \text{I} & x_1 & x_3 \\ \text{II} & x_2 & x_4 \end{array}$$

where players I and II play reals x_i .

I wins iff

- (a) there is an even i such that $\phi(x_i) = \infty$ and for all $j < i$, $\phi(x_j) < \infty$
- or
- (b) for all i , $\phi(x_i) < \infty$ and $\langle \phi(x_i) : i \in \omega \rangle \in A$.

Harrington–Kechris showed that Γ' having scales implies that this game is determined.

In practice we view players I and II as directly playing the ordinals $\phi(x_i)$. This seems to cause no functional difficulties.

MARTIN (1983) showed that the real game quantifier preserves scales. The real game quantifier applied to the least adequate class containing the scales corresponding to T_{\leq} and T_{\neq} gives us a class Γ^* with $\Gamma^* \supseteq \Gamma \supseteq \Gamma^*$. Hence every set in Γ has a scale.

Further facts we shall use without proof:

- (1) Γ is closed under \wedge , \vee , universal and existential real quantification (since κ is admissible).
- (2) there is a pairing function $\langle \ \rangle : \lambda \times \lambda \xrightarrow[\text{onto}]{1-1} \lambda$ that is in Δ .
- (3) Γ is normed and the norms take values in κ . (The norms are given by the order of constructibility of witnesses to the Σ_1 formulas).
- (4) The Boundedness theorem holds in the following guise:
Let $Y \in \Gamma$ and $\phi : Y \rightarrow \kappa$ be a Γ norm. Let $X \subseteq Y$.

- (1) if $X \in \check{\Gamma}$ then there is an $\alpha < \kappa$ $\phi''X \subseteq \alpha$.
- (2) if $\delta \in \lambda \cup Y$ and $\phi''X \subseteq \phi(\delta)$ then $X \in \check{\Gamma}$.

- (5) There are Γ -universal sets in Γ : there are $U \subseteq \lambda \times (\lambda \cup \mathbb{R})^n$, ($n \in \omega$), $U \in \Gamma$ and if $Y \subseteq (\lambda \cup \mathbb{R})^n$, $Y \in \Gamma$ then for some $\delta \in \lambda$, $Y = U_{\{\delta\}}$.
- (6) The fixed point theorem holds in the following form:

If $g : \lambda \rightarrow \lambda$ is in Γ then there is an α such that $U_{\{\alpha\}} = U_{\{g(\alpha)\}}$. Further, from a Γ index for g one can find the α in a uniformly Δ way.

- (7) There are sets $A \subseteq \lambda \times (\mathbb{R} \cup \lambda)$, $B \subseteq \lambda \times (\mathbb{R} \cup \lambda)$ and $C \subseteq \lambda$, $A, B, C \in \Gamma$ such that
 - (a) if $\alpha \in C$ then $B_{\{\alpha\}} = \mathbb{R} \setminus A_{\{\alpha\}}$
 - (b) if $X \in \Delta$ then for some $\alpha \in C$, $X = A_{\{\alpha\}}$.

We will call C the set of “ Δ -codes”.

(8) If $X \subseteq \{\Delta\text{-codes}\}$ is in Δ then there is an Δ -enumeration $\langle A_{\delta_i} : i < o.t.X \rangle$.

(9) Existential quantification over Δ is in Γ .
 (8, 9 follow immediately from 7)

2. The reflection lemmas

The facts cited above suffice to prove for Γ the facts about Π_1^1 proven in HARRINGTON *et al.* (unpublished). For the readers convenience we reproduce the proofs there mutatis mutandis using their notation and terminology with Γ replacing Π_1^1

DEFINITION 2.1. Suppose $A \subseteq \mathcal{P}(\mathbb{R})$. Then A is Γ on Γ iff $\{\alpha : U_{(\alpha)} \in A\}$ is in Γ .

LEMMA 2.2 (Reflection Lemma). Suppose A is Γ on Γ and $Y \in A$ is in Γ . Then there is an $X \in \Delta$, $X \subseteq Y$ and $X \in A$.

PROOF. Let $U \subseteq \lambda \times (\mathbb{R} \cup \lambda)$ be the Γ universal set.

Let $\psi : U \rightarrow \kappa$ be a Γ -norm. (We view $\psi : \lambda \times (\mathbb{R} \cup \lambda) \rightarrow \kappa \cup \{\infty\}$.) Let $\alpha, \beta \in \lambda$ be such that $U_{(\alpha)} = \{\gamma : U_{(\gamma)} \in A\}$ and $Y = U_{(\beta)}$.

For $\delta \in \lambda$, let $V_\delta = \{y \in Y : \psi(\beta, y) < \psi(\alpha, \delta)\}$. Then $V_\delta \in \Gamma$ and $V_\delta = U_{(\delta')}$ some δ' . Further, the map $\delta \mapsto \delta'$ is in Γ . By the fixed point theorem there is a δ such that $V_\delta = U_\delta$, and δ can be found in a uniformly Δ way from (α, β) .

If $U_\delta \notin A$ the $\psi(\alpha, \delta) = \infty$. But then $V_\delta = Y$ and $U_\delta = V_\delta$ so $U_\delta \in A$, a contradiction.

Hence, $U_\delta \in A$. Let $\gamma = \psi(\alpha, \delta)$. By fact 4 of the descriptive set theory summary $U_\delta = \{y \in Y : \psi(\beta, y) < \gamma\} \in \Delta$. \square

DEFINITION 2.3. $A \subseteq \mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ is Γ on Γ iff $\{(\alpha, \beta) : A(U_{(\alpha)}, U_{(\beta)})\}$ is in Γ .

A is *monotone upward* iff whenever $Y \subseteq Y', Z \subseteq Z'$ and $A(Y, Z)$ then $A(Y', Z')$. A is *continuous downward* iff whenever $Y_0 \supseteq Y_1 \supseteq \dots, Z_0 \supseteq Z_1 \supseteq \dots$, and $A(Y_i, Z_i)$ then $A(\bigcap_{i \in \omega} Y_i, \bigcap_{i \in \omega} Z_i)$.

LEMMA 2.4 (Strong Reflection). If $A \subseteq \mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ is Γ on Γ , monotone

upward, continuous downward and $Y \in \Gamma$, $A(Y, \sim Y)$ then there is an $X \in \Delta$ such that $X \subseteq Y$, $A(X, \sim X)$.

PROOF.

CLAIM. If $X \subseteq Y$, $X \in \Delta$ then there is a $\hat{X} \in \Delta$ $\hat{X} \supseteq X$ such that $A(\hat{X}, \neg X)$ and $\hat{X} \subseteq Y$.

PROOF OF CLAIM. Let $B(Z) = \{Z : A(Z, \neg X) \text{ and } X \subseteq Z\}$. Since A is Γ on Γ , B is Γ on Γ . Since A is monotone and $\sim X \supseteq \sim Y$, $A(Y, \sim X)$ so $B(Y)$. Hence by reflection, there is an $\hat{X} \subseteq Y$, $B(\hat{X})$. The claim follows. \square

We note that the use of the fixed point theorem in 2.2 was uniform, hence a Δ -code for \hat{X} can be gotten uniformly from a Δ -code for X and codes for Y , A .

Using the claim, build a Δ -sequence $\langle X_n : n \in \omega \rangle$ such that $Y \supseteq X_{n+1} \supseteq X_n$ and $A(X_{n+1}, \neg X_n)$. (Take $X_0 = \emptyset$.)

Let $X = \bigcup X_n$. Then $X \in \Delta$ and by monotonicity $A(X, \neg X_n)$ for each n . By continuity, $A(X, \bigcap \neg X_n)$. Hence, $A(X, \neg X)$.

An immediate consequence of strong reflection is:

COROLLARY 2.5. If $Y^* \in \tilde{\Gamma}$ is pre-linearly ordered by \leq then there is an $X \in \Delta$, $X \supseteq Y^*$ and X is pre-linearly ordered by \leq .

PROOF. Let $A(Y)$:

$$\forall x_0 x_1 \in \mathbb{R} (x_0 \notin Y \wedge x_1 \notin Y \Rightarrow x_0 \leq x_1 \vee x_1 \leq x_0)$$

Then A is Γ on Γ and $A(\neg Y^*)$. Hence there is an $X^* \subseteq \neg Y^*$ such that $X^* \in \Delta$ and $A(X^*)$. Let $X = \neg X^*$.

3. \leq -Separation

Following HARRINGTON *et al.* (unpublished) we now prove some lemmas about separation. These are analogous to Suslin's theorem.

Fix a thin \leq , λ_1 , λ_2 as in §1.

We need slightly different equivalence relations to prove the Dilworth theorem and the embedding theorem.

Let $\mathcal{F}_s = \{f: f \in \Delta \text{ and } f: \mathbb{R} \rightarrow 2^\alpha \text{ some } \alpha < \kappa \text{ and if } f(x) < f(y) \text{ then } x < y\}$. These are the “strongly order preserving functions”. Let $\mathcal{F}_w = \{f: f \in \Delta \text{ and } f: \mathbb{R} \rightarrow 2^\alpha \text{ some } \alpha < \kappa \text{ and if } x \leq y \text{ then } f(x) \leq f(y)\}$ (the “weakly order preserving functions”). Clearly \mathcal{F}_s and \mathcal{F}_w are in Γ .

If $Y \in \check{\Gamma}$ then we can also define $\mathcal{F}_s^Y = \{f: f \in \Delta \text{ and } f: \mathbb{R} \rightarrow 2^\alpha \text{ some } \alpha < \kappa \text{ and for all } x, y \in Y (f(x) < f(y) \Rightarrow x < y)\}$. Again \mathcal{F}_s^Y is in Γ .

For each of $\mathcal{F}_s, \mathcal{F}_w, \mathcal{F}_s^Y$, we get a $\check{\Gamma}$ equivalence relation E_s, E_w, E_Y by $xEy \leftrightarrow \forall f \in \mathcal{F}(f(x) = f(y))$.

Theorems 1.1 and 1.2 will follow by comparing these equivalence relations with the natural equivalence relation $x \sim y$ if $x \leq y$ and $y \leq x$.

LEMMA 3.1 (\leq -Separation). *Suppose $A, B \in \check{\Gamma} \cap \mathcal{P}(\mathbb{R})$ are disjoint and for all $a \in A, b \in B$ ($aEb \Rightarrow a \not\leq b$) then there is a $C \in \Delta, A \subseteq C$ and $B \cap C = \emptyset$ and for all $c \in C, d \notin C$ if cEd then $c \not\leq d$. (E is any of E_w, E_s or E_Y .)*

PROOF. (a) Let $A_0 = \{y: \exists x \in A(xEy \wedge y \leq x)\}$, $B_0 = \{y: \exists x \in B(xEy \wedge y \geq x)\}$. Then A_0, B_0 are disjoint and in $\check{\Gamma}$.

A_0 is downward closed in each E -equivalence class it intersects and B is upwards closed in each E -equivalence class it intersects.

Let $P(X, Y)$:

$$\forall xy[(xEy \wedge y \leq x) \Rightarrow (x \in X \vee y \in Y)] \wedge \forall x(x \in B_0 \rightarrow x \in X)$$

Then P is Γ on Γ , upwards monotone and continuous downward. Further, we have $P(\neg A_0, A_0)$. Hence, by strong reflection we get $P(\neg C, C)$ for some $C \in \Delta, \neg C \subseteq \neg A_0$. Hence $A \subseteq C, C \cap B = \emptyset$ and if $d \notin C, c \in C$ and cEd then $c \not\leq d$. □

LEMMA 3.2. *Suppose P is Γ or Γ and $P(\mathcal{F})$. Then for some $\mathcal{G} \subseteq \mathcal{F}, \mathcal{G} \in \Delta$ and $P(\mathcal{G})$. (Here again \mathcal{F} is any of $\mathcal{F}_w, \mathcal{F}_s, \mathcal{F}_s^Y$.)*

PROOF. This is an instance of the reflection lemma.

LEMMA 3.3 (Glue together Lemma). *Let $A, B \in \check{\Gamma} \cap \mathcal{P}(\mathbb{R}), A \cap B = \emptyset$.*

- (a) *If for all $a \in A, b \in B$ $aE_w b$ implies $a \not\leq b$ then for all $a \in A, b \in B$ $a \not E_w b$. (and similarly with $\not\leq$ replaced by $\not\leq$)*
- (b) *If $Y \supseteq A \cup B$ and there is a $C \in \Delta, A \subseteq C \subseteq \neg B$ and for all $c \in C \cap Y, d \in \check{C} \cap Y, cE_Y d$ implies $c < d$ then for all $a \in A, b \in B$ $a \not E_Y b$.*

PROOF. (a) By \leq -separation we can find a $C \in \Delta$ such that $a \subseteq C$, $C \cap B = \emptyset$ and C is \leq -downward closed in each E_w -equivalence class it intersects.

Let

$$P(X) \leftrightarrow \forall x, y \in \mathbb{R} [(\exists f: \mathbb{R} \rightarrow 2^\alpha \text{ some } \alpha, f \in X \text{ and } f(x) \neq f(y)) \vee (y \leq x \wedge x \in C \Rightarrow y \in C)].$$

Then $P(\mathcal{F}_w)$ so by 3.2 there is a $\mathcal{G} \subseteq \mathcal{F}_w$, \mathcal{G} in Δ with $P(\mathcal{G})$. By Δ -enumeration (property 8 of §1) we can enumerate \mathcal{G} in a Δ -way, $\langle f_\beta : \beta \in \gamma \rangle$, $\gamma \leq \lambda$. By boundedness there is an $\alpha < \kappa$ for all $\beta \in \gamma$, f_β takes values in 2^{α_β} some $\alpha_\beta < \alpha$.

Define $F^*: \mathbb{R} \rightarrow 2^{\sum_{\beta < \gamma} \alpha_\beta}$ by

$$F^*(x) \left[\left(\sum_{\beta < \beta'} \alpha_\beta \sum_{\beta < \beta'} \alpha_\beta + \alpha_{\beta'} \right) = f_{\beta'}(x) \right].$$

Then $F^* \in \Delta$ and if $x \leq y$ then $F^*(x) \leq F^*(y)$, so $F^* \in \mathcal{F}_w$.

Since $P(\mathcal{G})$, if $F^*(x) = F^*(y)$ and $x \in C$, $y \leq x$ then $y \in C$.

Let G be defined by

$$G(x) = \begin{cases} F^*(x) \frown 0 & \text{if } x \in C \\ F^*(x) \frown 1 & \text{if } x \notin C \end{cases}$$

Then G is order preserving. Further if $c \in C$, $d \notin C$ then G is a witness to $c \not E_w d$. Hence for all $a \in A$, $b \in B$ $a \not E_w b$.

(b) Using reflection on:

$$P(X) \leftrightarrow \forall x, y ((\exists f: \mathbb{R} \rightarrow 2^\alpha \text{ (some } \alpha), f \in X \text{ and } f(x) \neq f(y)) \vee (x, y \in Y \wedge x \in C \wedge y \notin C \Rightarrow x < y)),$$

there is a $\mathcal{G} \subseteq \mathcal{F}_Y$, $\mathcal{G} \in \Delta$ with $P(\mathcal{G})$.

As above there is an $F^*: \mathbb{R} \rightarrow 2^\alpha$ ($\alpha < \kappa$) $F^* \in \Delta$, F^* strongly order preserving and if $x, y \in Y$, $F^*(x) = F^*(y)$, $x \in C$ and $y \notin C$ then $x < y$.

Let

$$G(x) = \begin{cases} F^*(x) \frown 0 & x \in C \\ F^*(x) \frown 1 & x \notin C \end{cases}.$$

Then $G \in \mathcal{F}_Y$ and G witnesses that for all $a \in A$, $b \in B$ $a \not E_Y b$.

We note that the method of Lemma 3.3 gives a general way of glueing together a Δ -set of Δ functions.

4. The games

We will replace the forcing in HARRINGTON *et al.* (unpublished) by games. This is reasonable, as saying that σ is a winning strategy in a game is a claim that meeting certain dense sets guarantees some property.

We have two sorts of games (corresponding to \mathbb{P}_E^2 and $\mathbb{P} \times_E \mathbb{P}$ in HARRINGTON *et al.* (unpublished)).

Let $S \in \Gamma \cup \check{I}$, $A \in \check{I}$ and $A, S \subseteq \mathbb{R} \times \mathbb{R}$. Let E be an equivalence relation on \mathbb{R} . The game $G^2(A, E, S)$ is played as follows:

$$\begin{array}{ccccccc} \text{I} & B_0 & & B_2 & & \dots & \\ & & & & & & \\ \text{II} & & B_1 & & B_3 & & \dots \end{array}$$

I and II play \check{I} subsets B_i of A such that:

- (1) $B_{i+1} \subseteq B_i$
- (2) for all $(x, y) \in B_i$ xEy
- (3) diameter $(B_i) < (1/i + 1)$

II wins $G^2(A, E, S)$ iff $\bigcap_{i \in \omega} \overline{B_i} = (b_0, b_1) \in S$. (We will say in this case "II wins S ".)

We note that if $S \in \Gamma$ has a λ -Suslin representation in \check{I} and for all $B \in \check{I} \cap \mathcal{P}(A)$, $B \cap S \neq \emptyset$ then II has a winning strategy in $G^2(A, E, S)$. Similarly, if $\sim S \in \check{I}$ and has a λ -Suslin representation in \check{I} and $\sim S \cap A \neq \emptyset$ then I wins $G^2(A, E, S)$.

The other game we will be concerned with is $G(B, C, E, S)$ where $B, C \in \check{I} \cap \mathcal{P}(\mathbb{R})$, $S \in (\Gamma \cup \check{I}) \cap (\mathbb{R} \times \mathbb{R})$ and E is an equivalence relation.

In this game players I and II play pairs of \check{I} sets (B_i^0, B_i^1)

$$\begin{array}{ccccccc} \text{I} & (B_0^0, B_0^1) & & (B_2^0, B_2^1) & & \dots & \\ & & & & & & \\ \text{II} & & (B_1^0, B_1^1) & & (B_3^0, B_3^1) & & \dots \end{array}$$

such that

- (a) $B_{i+1}^0 \subseteq B_i^0 \subseteq B, B_{i+1}^1 \subseteq B_i^1 \subseteq C$
- (b) for all i there are $b_i^0 \in B_i^0, b_i^1 \in B_i^1$, with $b_i^0 E b_i^1$
- (c) $\text{diam}(B_i^j) < (1/i + 1)$

Player II wins iff

$$\left(\bigcap_{i \in \omega} \overline{B_i^0}, \bigcap_{i \in \omega} \overline{B_i^1} \right) = (b_0, b_1) \in S.$$

DEFINITION 4.1. Let \mathcal{S} be a strategy for either player I or II in either $G^2(A, E, S)$ or $G(B, C, E, S)$. A pair of reals (b_0, b_1) is \mathcal{S} -generic iff there is a play of the appropriate game by the strategy \mathcal{S} that produces (b_0, b_1) .

Some general remarks about these games are in order. First, if II has a winning strategy in $G(A, B, E, S)$ (or $G^2(D, E, T)$) then I has a winning strategy in $G(A, B, E, \sim S)$ (or $G^2(D, E, \sim T)$) simply by playing II's strategy against the trivial first move by I in an auxiliary game.

Secondly, if $\mathcal{S}_1, \mathcal{S}_2$ are two strategies for II in $G(A, B, E, S)$ (or $G^2(D, E, S)$) then II can meld them into one strategy \mathcal{S} by playing \mathcal{S}_2 against \mathcal{S}_1 's response to moves by I. In this way any \mathcal{S} -generic pair is both \mathcal{S}_1 and \mathcal{S}_2 -generic. The same is true if \mathcal{S}_1 and \mathcal{S}_2 are winning strategies for I provided that $\mathcal{S}_1(\phi)$ and $\mathcal{S}_2(\phi)$ are compatible.

Finally, since $\tilde{\Gamma}$ is indexed by ordinals in λ (via $\neg U$) we can view each of these games as games played on λ . Hence by the remarks in §2 on determinacy we may assume that all of these games are determined.

We now need a lemma which has a slightly more complicated proof than the forcing analogue.

LEMMA 4.2. Let $\langle E_A : A \in \tilde{\Gamma} \rangle$ be a sequence of $\tilde{\Gamma}$ equivalence relations such that if $A \subseteq A'$ then E_A refines $E_{A'}$. If $\leq \in \Gamma \cap \tilde{\Gamma}$, is a thin quasi-ordering then for all $B \in \tilde{\Gamma}$ there is a (non-empty) $B' \in \tilde{\Gamma} \cap \mathcal{P}(B)$ for all $A \in \tilde{\Gamma} \cap \mathcal{P}(B')$ II wins $G(A, A, E_A, \leq\text{-comparable})$ (i.e. $S = \{(x, y) : x \text{ is } \leq\text{-comparable to } y\}$).

Note: If II wins $G(A, A, E_A, \leq\text{-comparable})$ and $A' \in \tilde{\Gamma} \cap \mathcal{P}(A)$ then II wins $G(A', A', E_A, \leq\text{-comparable})$.

NOTATION. We will write \approx for comparable and $\not\approx$ for incomparable.

PROOF. Otherwise we can find a $B \in \check{I}$ such that for all $B' \in \check{I} \cap \mathcal{P}(B)$ there is an $A \in \check{I} \cap \mathcal{P}(B')$ such that I wins $G(A, A, E_A, \approx)$. If I wins $G(A, A, E_A, \approx)$ then A is not a \check{I} -singleton. Hence B contains no \check{I} -singletons and each $B' \in \check{I} \cup \mathcal{P}(B)$ can be split into two disjoint \check{I} subsets.

Let $\text{Seq} = \{s : \exists n \in \omega, s : n \rightarrow 2\}$

CLAIM. There are \check{I} sets $\langle C_s : s \in \text{Seq} \rangle, \langle A_s : s \in \text{Seq} \rangle$ such that:

- (a) for all $s \in \text{Seq}$, I wins $G(A_s, A_s, E_s, \approx)$ where $E_s = E_{A_s}$.
- (b) If $s'00 \supseteq s$ then $A_{s'} \subsetneq C_{s'} \subsetneq A_s \subsetneq C_s$.
- (c) If \mathcal{S}_s is I's winning strategy in $G(A_s, A_s, E_s, \approx)$ and $(B_0^0, B_0^1) = \mathcal{S}_s(\phi)$ then $C_{s \cdot 0} \subseteq B_0^0$ and $C_{s \cdot 1} \subseteq B_0^1$.
- (d) $\text{diam}(C_s) < (1/\ell(s) + 1)$.
- (e) For each $t : m \rightarrow 2$ and each $s, s' : n \rightarrow 2$ with $s \wedge s' = t, s >_{\text{lex}} s'$ we can choose $\langle (B_i^0, B_i^1) : i \in 2(n-m) - 1 \rangle_{ss'}$ a partial play of $G(A_t, A_t, E_t, \approx)$ according to \mathcal{S}_t such that $C_s \subset B_{2(n-m)-2}^0$ and $C_{s'} \subseteq B_{2(n-m)-2}^1$.
- (f) If r, r' extend s, s' then $\langle (B_i^0, B_i^1) : i \in 2(\ell(r) - m) - 1 \rangle_{rr'}$ extends $\langle (B_i^0, B_i^1) : i \in 2(n-m) - 1 \rangle_{ss'}$.
- (g) For all $x \in C_s$ there are $\langle y_{s'} : \ell(s') = \ell(s) \rangle$ with $y_s = x$ such that $y_{s'} \in C_{s'}$ and if $t = s_1 \wedge s_2$ then $y_{s_1} E_t y_{s_2}$.

To see that the claim suffices: for each $f \in {}^\omega 2$ let $b_f = \bigcap_{n \in \omega} \bar{C}_{f \upharpoonright n}$. The set $\{b_f : f \in {}^\omega 2\}$ is a perfect set.

Let $f_0 <_{\text{lex}} f_1$. Let $t = f_0 \wedge f_1$. Then there is a play of the game $G(A_t, A_t, E_t, \approx)$ by \mathcal{S}_t given by

$$\{ \langle (B_i^0, B_i^1) : i \in 2(n - \ell(t)) - 1 \rangle_{f_0 \upharpoonright n, f_1 \upharpoonright n} : n \in \omega \}.$$

By $f, \bigcap_{n \in \omega} \bar{C}_{f \upharpoonright n} = \bigcap_{n \in \omega} \bar{B}_{2(n-\ell(t))}^i$ for $i = 0, 1$. Hence, (b_{f_0}, b_{f_1}) is \mathcal{S}_t generic and thus $b_{f_1} \not\leq b_{f_0}$. This shows that there are a perfect set of \leq -incomparable reals, a contradiction.

PROOF OF CLAIM. We build $\langle A_s : s \in \text{Seq} \rangle \langle C_s : s \in \text{Seq} \rangle$ and $\langle (B_i^0, B_i^1) \rangle_{ss'}$ by induction on $\ell(s) =$ the length of s .

Suppose we have succeeded in constructing C 's and $\langle (B_i^0, B_i^1) \rangle$'s satisfying (a)–(g) for all s with $\ell(s) \leq n$ and the A 's for s with $\ell(s) \leq n - 1$. Let $\langle s_i : i \in 2^n - 2^{n-1} \rangle$ enumerate the sequences of length n .

By induction on $i \in 2^n - 2^{n-1}$ we choose $A_{s_i}, \text{diam}(A_{s_i}) < (1/n + 2)$ and

sets $D_t^i \in \tilde{I}$, $t \in {}^{n+1}2$ such that for all i and all $t \in {}^{n+1}2$ and all $x \in D_t^i$, there are $\{y_{t'} : t' \in {}^{n+1}2\}$, $y_t = x$ and for all $t', t'' \in {}^{n+1}2$ if $t^* = t' \wedge t''$ then $y_{t'} E_{t'} y_{t''}$ (i.e. the $\{D_t^i : t \in {}^{n+1}2\}$ satisfy condition (g) in the claim). Also if $j > i$ then $D_{s_{j-0}}^i = D_{s_{j-1}}^i$.

Suppose, we are at stage i in the induction. Then $D_{s_{i-0}}^{i-1} = D_{s_{i-1}}^{i-1}$. Choose $A_{s_i} \subseteq D_{s_{i-0}}^{i-1}$ such that $A_{s_i} \in \tilde{I}$, $\text{diam}(A_{s_i}) < (1/n + 2)$ and I wins $G(A_{s_i}, A_{s_i}, E_{s_i}, \approx)$.

Let \mathcal{S}_{s_i} be a winning strategy for I in this game. Let $(B_0^0, B_1^0) = \mathcal{S}_{s_i}(\emptyset)$. Let $D_{s_{i-0}}^i \subseteq B_0$ and $D_{s_{i-1}}^i \subseteq B_1$ be so that for all $x \in D_{s_{i-0}}^i$ there is a $y \in D_{s_{i-1}}^i$ $x E_{s_i} y$ and vice versa.

For each $j \neq i$, and $k = 0$ or 1 , let $D_{s_{j-k}}^i = \{x \in D_{s_{j-k}}^{i-1} : \text{there is a sequence } \langle y_t : t \in {}^{n+1}2 \rangle \text{ with } y_t \in D_t^{i-1} \text{ and } y_{s_{j-k}} = x \text{ and } y_{s_{i-0}} \in D_{s_{i-0}}^i \text{ and } y_{s_{i-1}} \in D_{s_{i-1}}^i \text{ such that for all } t, t' \subseteq {}^{n+1}2 \text{ if } t^* = t \wedge t' \text{ then } y_t E_{t'} y_{t'}\}$.

Since each $E_{t'}$ is an equivalence relation $D_{s_{j-k}}^i \neq \emptyset$. Since each $E_{t'}$ is \tilde{I} we get that $D_{s_{j-k}}^i \in \tilde{I}$. Again, using the fact that $E_{t'}$ is an equivalence relation the $\langle D_t^i : t \in {}^{n+1}2 \rangle$ satisfy the induction hypothesis for the D 's.

For each $t \in {}^{n+1}2$, let $X_t^0 = D_t^{2^n - 2^{(n-1)}}$. For each $s \in {}^n 2$, let $\langle (B_0^0, B_1^0) \rangle_{s_{-0}, s_{-1}} = \mathcal{S}_s(\emptyset)$.

Let $\{(t_1, t_2) : i \in i^*\}$ enumerate all pairs of functions $t_1, t_2 : n + 1 \rightarrow 2$ with $t_1 <_{\text{lex}} t_2$. We now build the plays $\langle (B_i^0, B_i^1) : i \in 2(n + 1 - \ell(t_1 \wedge t_2)) - 1 \rangle_{t_1 t_2}$.

We will build by induction a sequence $\{X_t^i \mid t : n + 1 \rightarrow 2 \text{ and } i \in i^*\}$. If $i < i'$ then $X_t^i \supseteq X_t^{i'}$. For each i , $\{X_t^i\}$ will satisfy (g) of the claim, i.e. if $x \in X_t^i$ then there is a sequence $y_{t'} \in X_{t'}^i$ ($t' : n + 1 \rightarrow 2$) with $y_{t'} = x$ and for all $t^*, t' : n + 1 \rightarrow 2$ if $s^* = t^* \wedge t'$ then $y_{t'} E_{s^*} y_{t'}$.

At stage i we want to build the X_t^{i+1} 's. Consider $(t_1, t_2)_i$. Let s_1 and s_2 be such that $t_1 = s_1 k_1$, $t_2 = s_2 k_2$.

If $s_1 = s_2$ then we have already defined $\langle (B_0^0, B_0^1) \rangle_{t_1 t_2}$. For all t , let $X_t^{i+1} = X_t^i$.

If $s_1 \neq s_2$, let $s^* = s_1 \wedge s_2 = t_1 \wedge t_2$. Since $s_1 <_{\text{lex}} s_2$ we have defined a play of $G(A_{s^*}, A_{s^*}, E_{s^*}, \approx)$ according to \mathcal{S}_{s^*} , $\langle (B_i^0, B_i^1) : i \in 2(n - \ell(s^*)) - 1 \rangle$. By (e) in the claim and the induction hypothesis on C_{s_1} and C_{s_2} we have that $X_{t_1}^i \subseteq C_{s_1} \subseteq B_{2(n-\ell(s^*))-2}^0$ and $X_{t_2}^i \subseteq C_{s_2} \subseteq B_{2(n-\ell(s^*))-2}^1$. Since for all $x \in X_{t_1}^i$ there is a $y \in X_{t_2}^i$, $x E_{s^*} y$ and vice versa, $(X_{t_1}^i, X_{t_2}^i)$ is a legal next move in $G(A_{s^*}, A_{s^*}, E_{s^*}, \approx)$. (This is what all of the fuss is about). Let $B_{2(n-\ell(s^*))-1}^0 = X_{t_1}^i$ and $B_{2(n-\ell(s^*))-1}^1 = X_{t_2}^i$. Let

$$(B_{2(n-\ell(s^*))}^0, B_{2(n-\ell(s^*))}^1) =$$

$$\mathcal{S}_{s^*}(\langle (B_i^0, B_i^1) : i \in 2(n - \ell(s^*)) - 2 \rangle_{s_1 s_2} (B_{2(n-\ell(s^*))-1}^0, B_{2(n-\ell(s^*))-1}^1)) .$$

Let

$$\begin{aligned} & \langle (B_i^0, B_i^1) : i \in 2(n+1 - \ell(s^*)) - 1 \rangle_{t_1 t_2} \\ &= \langle (B_i^0, B_i^1) : i \in 2(n - \ell(s^*)) - 1 \rangle_{s_1 s_2} \\ & \quad \langle (B_{2(n-\ell(s^*))-1}^0, B_{2(n-\ell(s^*))-1}^1), (B_{2(n-\ell(s^*))}^0, B_{2(n-\ell(s^*))}^1) \rangle \end{aligned}$$

Then $\langle (B_i^0, B_i^1) : i \in 2(n+1 - \ell(s^*)) - 1 \rangle_{t_1 t_2}$ is a legal partial play of the game $G(A_{s^*}, A_{s^*}, E_{s^*}, \infty)$ according to \mathcal{S}_{s^*} and extends the plays corresponding to all $r_1 \subseteq t_1$ and $r_2 \subseteq t_2$ previously defined.

Now let $X_{t_1}^{i+1} \subseteq B_{2(n-\ell(s^*))}^0$ and $X_{t_2}^{i+1} \subseteq B_{2(n-\ell(s^*))}^0$ be such that for all $x \in X_{t_1}^{i+1}$ there is a $y \in X_{t_2}^{i+1}$ $x E_{s^*} y$ and vice versa.

For $t_0 \neq t_1, t_2$ let $X_{t_0}^{i+1} = \{x \in X_{t_0}^i : \text{there is a sequence } \langle y_t | t : n+1 \rightarrow 2 \rangle,$

- (1) $y_t \in X_t^i$ for all t
- (2) $y_{t_1} \in X_{t_1}^{i+1}, y_{t_2} \in X_{t_2}^{i+1}, y_{t_0} = x$
- (3) for all $t, t' \in \overset{n+1}{2}, y_t E_{t \wedge t'} y_{t'}$.

Then each X_t^{i+1} is in $\check{\Gamma}$. We must argue that each $X_t^{i+1} \neq \emptyset$.

To show this it is enough to find a single sequence $\langle y_t | t : n+1 \rightarrow 2 \rangle$ with $y_t \in X_t^i, y_{t_1} \in X_{t_1}^{i+1}$ and $y_{t_2} \in X_{t_2}^{i+1}$ such that for all $t, t' : n+1 \rightarrow 2, y_t E_{t \wedge t'} y_{t'}$.

Choose $y_{t_1} \in X_{t_1}^{i+1}$ and $y_{t_2} \in X_{t_2}^{i+1}, y_{t_1} E_{s^*} y_{t_2}$. By the induction hypothesis on the X_t^i 's there is a sequence $\langle x_t | t : n+1 \rightarrow 2 \rangle$ with $x_t \in X_t^i$, with $x_{t_1} = y_{t_1}$ and $x_t E_{t \wedge t'} x_{t'}$, and a sequence $\langle z_t | t : n+1 \rightarrow 2 \rangle$ with $z_t \in X_t^i, z_{t_2} = y_{t_2}$ and $z_t E_{t \wedge t'} z_{t'}$.

CASE 1. $t : n+1 \rightarrow 2$ and $t \wedge t_1 \supseteq s^*$. Let $y_t = x_t$.

CASE 2. $t : n+1 \rightarrow 2$ and $t \wedge t_1 = s^*$. Let $y_t = z_t$.

CASE 3. Otherwise $y_t = x_t$.

Let $t, t' \in \overset{n+1}{2}$. We need to see $y_t E_{t \wedge t'} y_{t'}$.

If $t \wedge t' \supseteq s^*$, then either $y_t = x_t$ and $y_{t'} = x_{t'}$, or $y_t = z_t$, $y_{t'} = z_{t'}$. In either case there is no problem by the choice of x_t and z_t 's.

If $t \wedge t' = s^*$ and $t <_{\text{lex}} t'$ then $y_t E_{s^*} y_{t_1}$, since $E_{t \wedge t_1}$ refines E_{s^*} . Similarly, $y_{t'} E_{s^*} y_{t_2}$. Since $y_{t_1} E_{s^*} y_{t_2}$, we have $y_t E_{s^*} y_{t'}$.

If $t \wedge t' \subsetneq s^*$, then either $y_t = x_t$ and $y_{t'} = x_{t'}$ (so there is no difficulty) or either $y_t = z_t$ or $y_{t'} = z_{t'}$. Assume $y_t = z_t$ and $y_{t'} = x_{t'}$. Then $z_t E_{s^*} y_{t_1}$

and E_s refines $E_{t \wedge t'}$ and $y_t E_{t \wedge t'} x_{t'}$ so $y_t E_{t \wedge t'} y_{t'}$. In the case that $t \wedge t'$ is \subseteq incomparable with s^* then $y_t = x_t$ and $y_{t'} = x_{t'}$ so there is no problem.

Hence we have constructed $\langle X_i^i | t: n + 1 \rightarrow 2, i < i^* \rangle$. Let $C_i = X_i^{i^*-1}$. Then $\langle C_i | i \in \omega \rangle$ satisfy (a)–(g) in the claim for sequences of length $n + 1$. This completes the inductive construction. \square

REMARK. What prevents an entirely routine dove-tailing argument for the lemma is the problem that if $t \not\subseteq t'$ the strategy \mathcal{S}_i might call for an illegal play in $G(A_{t'}, A_{t'}, E_{t'}, \cong)$. To prevent this we only present \mathcal{S}_i with positions where it can not do this, i.e., condition (g).

By Lemma 4.2, on a dense open set, player II has a strategy to “force” two reals to be \leq -comparable.

We will use II’s strategy to get contradictions to the transitivity of \leq (“order contradictions”) but we need to be able to dove-tail II’s strategy with other strategies to get “mutually generic” reals. This is the thrust of the next Lemma.

LEMMA 4.3. *Let (A, B) be in $\check{\Gamma}$, E a $\check{\Gamma}$ equivalence relation with $(A \times B) \cap E \neq \emptyset$. Suppose $T, S \subseteq \mathbb{R}^2$ and I has a winning strategy \mathcal{S} for $G(A, B, E, S)$. Let $(A_0, B_0) = \mathcal{S}(\emptyset)$ and $D \in \check{\Gamma}$, $D \subseteq A_0 \times B_0$. Suppose II has a winning strategy \mathcal{T} for $G^2(D, E, T)$. Then there are reals $\{a_{ij} : i \in 3, j \in 2\}$ such that for each i (a_{i0}, a_{i1}) is \mathcal{T} -generic and if $i' \neq i$ then $(a_{i'0}, a_{i'1})$ is \mathcal{S} -generic.*

PROOF. As in the previous Lemma, the differing rules of G^2 and G prevent the completely straightforward construction.

We build sequences of $\check{\Gamma}$ sets $\langle \mathcal{A}_k^i : k \in \omega, i = 0, 1, 2 \rangle$ and $\langle A_k^{i,i'} : k \in \omega, i \neq i' \in 3 \rangle$ and $\langle B_k^{i,i'} : k \in \omega, i \neq i' \in 3 \rangle$. They will satisfy:

- (1) $\langle \mathcal{A}_k^i : k \in \omega \rangle$ is a play of $G^2(D, E, T)$ according to \mathcal{T}
- (2) for $i \neq i'$, $\langle (A_k^{i,i'}, B_k^{i,i'}) : k \in \omega \rangle$ is a play by \mathcal{S} in $G(A, B, E, S)$
- (3) If $i' < i''$ then

$$\text{for odd } k \begin{cases} A_{k+2}^{i,i'} \subseteq A_k^{i,i''} \subseteq A_k^{i,i'} \text{ and} \\ B_{k+2}^{i,i'} \subseteq B_k^{i,i''} \subseteq B_k^{i,i'} \end{cases}$$

- (4) for each i, i' and for all even k , $\pi_0(\mathcal{A}_k^i) \subseteq A_k^{i,i'}$ and $\pi_1(\mathcal{A}_k^i) \subseteq B_k^{i,i'}$.

(5) If k is even then for all $i, i'x \in \pi_0(\mathcal{A}_k^i)$ there is a $y \in \pi_1(\mathcal{A}_k^{i'})$ such that xEy and similarly with π_1 and π_0 reversed.

To see that this suffices let the \mathcal{A} 's, A 's and B 's satisfy (1)–(5).

By (3) $\bigcap_k \overline{A}_k^{i,i'} = \bigcap_k \overline{A}_k^{i,i''}$ and $\bigcap_k \overline{B}_k^{i,i'} = \bigcap_k \overline{B}_k^{i,i''}$ for all i, i', i'' .

By (4) $\bigcap_k \overline{\mathcal{A}}_k^i = (\bigcap_k \overline{A}_k^{i,i'}, \bigcap_k \overline{B}_k^{i,i'})$. Let $(a_{i,0}, a_{i,1}) = \bigcap_k \overline{\mathcal{A}}_k^i$. By (1), $(a_{i,0}, a_{i,1})$ is \mathcal{F} -generic. By (2) $(a_{i,0}, a_{i,1})$ is \mathcal{S} -generic if $i \neq i'$.

We perform this construction by induction on k . For each i, i' let $(A_0^{i,i'}, B_0^{i,i'}) = \mathcal{S}(\emptyset)$ and $\mathcal{A}_0^i = D$.

Suppose $k = 2j$ and we have constructed $\langle \mathcal{A}_{k'}^i : k' \leq k \rangle$ and $\langle A_{k'}^{i,i'} : k' \leq k \rangle$ and $\langle B_{k'}^{i,i'} : k' \leq k \rangle$ satisfying (1)–(5).

Let $\mathcal{A}_{k+1}^i = \mathcal{F}(\mathcal{A}_k^i)$. We begin building $A_{k+1}^{0,i}, A_{k+2}^{0,i}$ and $B_{k+1}^{i,0}, B_{k+2}^{i,0}$, $i = 1, 2$.

Let $A_{k+1}^{0,1} = \pi_0(\mathcal{A}_{k+1}^0)$ and $B_{k+1}^{1,0} = \pi_1(\mathcal{A}_{k+1}^1)$. By (4), (5) this is a legal move in $G(A, B, E, S)$ below $\langle (A_{k'}^{0,1}, B_{k'}^{1,0}) : k' \leq k \rangle$. Let $(A_{k+2}^{0,1}, B_{k+2}^{1,0}) = \mathcal{S}(\langle (A_{k'}^{0,1}, B_{k'}^{1,0}) : k' \leq k+1 \rangle)$. Let $A_{k+1}^{0,2} \subseteq A_{k+2}^{0,1}$ be such that for all $x \in A_{k+1}^{0,2}$ there is a $y \in B_{k+2}^{1,0}$, xEy . Let $B_{k+1}^{2,0} = \pi_1(\mathcal{A}_{k+1}^2)$. Then $(\mathcal{A}_{k+1}^{0,2}, B_{k+1}^{2,0})$ is a legal move in $G(A, B, E, S)$. Let $(A_{k+2}^{0,2}, B_{k+2}^{2,0}) = \mathcal{S}(\langle (A_{k'}^{0,2}, B_{k'}^{2,0}) : k' \leq k+1 \rangle)$. Let $A^*, B^*, B^{**} \in \tilde{\Gamma}$ be such that $A^* \subseteq A_{k+2}^{0,2}$, $B^* \subseteq B_{k+2}^{1,0}$, $B^{**} \subseteq B_{k+2}^{2,0}$ and for all $x \in A^*$ there are $y_1 \in B^*$, $y_2 \in B^{**}$ such that xEy_1Ey_2 and similarly for all y_1 there are x_2, y_2 etc. Let $\mathcal{C}_0^0 = \mathcal{A}_{k+1}^0 \cap \pi_0^{-1}(A^*)$, $\mathcal{C}_0^1 = \mathcal{A}_{k+1}^1 \cap \pi_1^{-1}(B^*)$ and $\mathcal{C}_0^2 = \mathcal{A}_{k+1}^2 \cap \pi_1^{-1}(B^{**})$. Then for all i, i' and all $x \in \pi_0(\mathcal{C}_0^i)$ there is a $y \in \pi_1(\mathcal{C}_0^{i'})$ with xEy and for all $y \in \pi_1(\mathcal{C}_0^{i'})$ there is $x \in \pi_0(\mathcal{C}_0^i)$, xEy .

Build $A_{k+1}^{1,0}, A_{k+2}^{1,0}, A_{k+1}^{1,2}, A_{k+2}^{1,2}$ and $B_{k+1}^{0,2}, B_{k+2}^{0,1}, B_{k+1}^{2,1}, B_{k+2}^{2,1}$ analogously by letting $A_{k+1}^{1,0} = \pi_0(\mathcal{C}_0^1)$ and $B_{k+1}^{0,1} = \pi_1(\mathcal{C}_0^0)$, $B_{k+1}^{2,1} = \pi_1(\mathcal{C}_0^2)$ and continuing as above. Let $A^* \subseteq A_{k+2}^{1,2}$, $B^* \subseteq B_{k+2}^{0,1}$ and $B^{**} \subseteq B_{k+2}^{2,1}$ be such that for all $x \in A^*$ there are $y_1 \in B^*$, $y_2 \in B^{**}$ xEy_1Ey_2 and for all $y_1 \in B^*$ there are $x \in A^*$, $y_2 \in B^{**}$ xEy_1Ey_2 , etc.

$\mathcal{C}_1^0 = \mathcal{C}_0^0 \cap \pi_1^{-1}(B^*)$, $\mathcal{C}_1^1 = \mathcal{C}_0^1 \cap \pi_0^{-1}(A^*)$ and $\mathcal{C}_1^2 = \mathcal{C}_0^2 \cap \pi_1^{-1}(B^{**})$. Then for all $i \neq i'$ and all $x \in \pi_0(\mathcal{C}_1^i)$ there is a $y \in \pi_1(\mathcal{C}_1^{i'})$ xEy and for all $y \in \pi_1(\mathcal{C}_1^{i'})$ there is an $x \in \pi_2(\mathcal{C}_1^i)$ xEy .

Build $A_{k+1}^{2,0}, A_{k+2}^{2,0}, A_{k+1}^{2,1}, A_{k+2}^{2,1}, B_{k+1}^{0,2}, B_{k+2}^{0,2}, B_{k+1}^{1,2}, B_{k+2}^{1,2}$ in the same way. Choose $A^* \subseteq A_{k+2}^{2,1}$, $B^* \subseteq B_{k+2}^{0,2}$, $B^{**} \subseteq B_{k+2}^{1,2}$ as before. Let $\mathcal{A}_{k+2}^0 = \mathcal{C}_1^0 \cap \pi_1^{-1}(B^*)$, $\mathcal{A}_{k+2}^1 = \mathcal{C}_1^1 \cap \pi_1^{-1}(B^{**})$ and $\mathcal{A}_{k+2}^2 = \mathcal{C}_1^2 \cap \pi_0^{-1}(A^*)$.

This completes the construction.

LEMMA 4.4. *Let $A, B \in \tilde{\Gamma}$. Suppose I has a winning strategy \mathcal{S}_1 in $G(A, B, E, S)$ and $\mathcal{S}_1(\emptyset) = (A_0, B_0)$. Let $A^* \in \tilde{\Gamma}$, $A^* \subseteq \{a \in A_0 \mid \exists b \in B_0 aEb\}$ and suppose I has a winning strategy \mathcal{S}_2 in $G(A^*, A^*, E, S_2)$. Let*

$\mathcal{S}_2(\emptyset) = (A', B')$ and let $D \subseteq (B' \times B_0) \cap E$ be such that II wins $G^2(D, E, T)$ by a strategy \mathcal{F} . Then there are reals a_0, a_1, b such that

- (a) (a_0, b) is \mathcal{S}_1 -generic
- (b) (a_0, a_1) is \mathcal{S}_2 -generic
- (c) (a_1, b) is \mathcal{T}_1 -generic

(and similarly if $D \subseteq (A' \times B_0) \cap E$).

PROOF. Proceed similarly to Lemma 4.3 to simultaneously build interlaced plays of $G(A, B, E, S_1)$, $G(A^*, A^*, E, S_2)$ and $G^2(D, E, T)$. By always restricting the plays by II in $G(A, B, E, S_1)$ and $G(A^*, A^*, E, S_2)$ to pairs (C, D) such that for all $c \in C$ there is $d \in D$ cEd and all $d \in D$ there is $c \in C$ cEd , the responses by I according to \mathcal{S}_1 and \mathcal{S}_2 are always legal in the various games. The details are left to the reader. \square

5. The embedding theorem

We prove Theorem 1.2.

Recall from §3 the definition of \mathcal{F}_w and E_w . For this section let $\mathcal{F} = \mathcal{F}_w$ and $E = E_w$.

We let $x \sim y$ iff $x \leq y$ and $y \leq x$. We will show:

MAIN CLAIM:

$$xE_w y \text{ iff } x \sim y.$$

To see the main claim suffices to show Theorem 1.2, we note that:

$$P(X) \leftrightarrow \forall xy \in \mathbb{R} [(\exists \alpha \exists f: \mathbb{R} \rightarrow 2^\alpha, f \in X \text{ and } f(x) \neq f(y)) \vee (x \leq y \wedge y \leq x)]$$

is Γ on Γ . Further $P(\mathcal{F})$. Hence, by Lemma 3.2 there is a $\mathcal{G} \in \Delta$, $\mathcal{G} \subseteq \mathcal{F}$ with $P(\mathcal{G})$. By Δ -enumeration we can enumerate \mathcal{G} as $\langle f_\beta: \beta \in \gamma \rangle$, $\gamma \leq \lambda$. Following the argument in the glue-together Lemma 3.3, if $f_\beta: \mathbb{R} \rightarrow 2^{\alpha_\beta}$ then $\delta = \sum_{\beta \in \gamma} \alpha_\beta < \kappa$ and $F^*: \mathbb{R} \rightarrow 2^\delta$ given by

$$F^*(x) \left[\left[\sum_{\beta < \beta'} \alpha_\beta, \sum_{\beta < \beta'} \alpha_\beta + \alpha_{\beta'} \right] \right] = f_{\beta'}(x)$$

is an order preserving Δ -function from \mathbb{R} into 2^δ , $\delta < \kappa$. By the main claim, if $F^*(x) = F^*(y)$ then $x \sim y$.

LEMMA 5.1. *Let $X \in \check{I}$ and suppose for all $x, y \in X(xEy \rightarrow x \sim y)$ then for all $x \in X$ and all $y (xEy \rightarrow x \sim y)$.*

PROOF. Let $A = \{y: \exists x \in X(xEy \wedge y \leq x)\}$, $B = \{y: \exists x \in X(xEy \wedge (x \not\approx y \vee y > x))\}$. We note $A \cup B = \{y: \exists x \in X(xEy)\}$.

CASE 1. $B \neq \emptyset$.

Then $A \cap B = \emptyset$ since otherwise, there are y, x_1, x_2 with $x_1, x_2 \in X, yEx_1, yEx_2$ and $y \leq x_1$ and $(y \approx x_2$ or $y > x_2)$. But then x_1Ex_2 and either $x_1 \approx x_2$ or $x_2 < x_1$ contradicting the hypothesis.

Further, if $a \in A, b \in B$ then $a \not\approx b$. Hence, by the glue-together lemma, for all $a \in A, b \in B, a \not\approx b$. But $X \subseteq A$ and $B \neq \emptyset$; a contradiction.

CASE 2. Otherwise. Let $A' = \{y: \exists x \in X(xEy \wedge y < x)\}$ and $B' = \{y: \exists x \in X(xEy \wedge y \geq x)\}$. Since $B = \emptyset, A' \cup B' = \{y: \exists x \in X(xEy)\}$. Again, $A' \cap B' = \emptyset$ and for all $a \in A', b \in B', a \not\approx b$. So by the glue together lemma, for all $a \in A', b \in B', a \not\approx b$. Since $X \subseteq B', A' = \emptyset$.

Since $B = \emptyset$, for all y and all $x \in X(xEy \rightarrow y \leq x)$. Since $A' = \emptyset$, for all y and all $x \in X(xEy \rightarrow x \leq y)$ \square .

Let $Z = \{x: \exists y(xEy \wedge x \not\approx y)\}$. Then $Z \in \check{I}$. If $Z = \emptyset$ then we are done.

CLAIM 1. For all $Y \in \check{I} \cap \mathcal{P}(Z)$ II has a winning strategy in $G(Y, Y, E, \not\approx)$ and in $G(Y, Y, E, \not\prec)$.

PROOF. Assume I wins $G(Y, Y, E, \not\prec)$ by strategy \mathcal{S} . Let (A, B) be I's first move. By the glue-together Lemma 3.3 there are $a \in A, b \in B(aEb \wedge a \geq b)$. Let $D = \{(a, b): a \in A, b \in B, aEb \wedge a \geq b\}$. Since \leq is λ -Suslin in \check{I} , II wins $G^2(D, E, \leq)$, say by \mathcal{T} .

By Lemma 4.3 there are $(a_0, a_1), (b_0, b_1)$ such that $(a_0, a_1), (b_0, b_1)$ are \mathcal{T} -generic and (a_0, b_1) and (b_0, a_1) are \mathcal{S} -generic. Hence, $a_0 \geq a_1, b_0 \geq b_1$ and $a_0 < b_1$ and $b_0 < a_1$. But then $b_0 < a_1 \leq a_0 < b_1 \leq b_0$, a contradiction. Hence II wins $G(Y, Y, E, \not\prec)$. The proof for $G(Y, Y, E, \not\approx)$ is similar. \square

CLAIM 2. For all $Y \subseteq Z$, II has a winning strategy in $G(Y, Y, E, \not\prec)$.

PROOF. Otherwise, let \mathcal{S} be I's strategy, and (A_0, B_0) be I's first move.

CASE 1. There are $a \in A_0, b \in B_0$ with aEb and $a < b$.

Let $D = \{(a, b) : aEb, a < b \text{ and } a \in A_0, b \in B_0\}$. Then II wins $G^2(D, E, <)$, since it is λ -Suslin. Let \mathcal{T} be II's strategy.

Let $(a_0, b_0), (a_1, b_1)$ be such that $(a_0, b_0), (a_1, b_1)$ are \mathcal{T} -generic and (a_0, b_1) and (a_1, b_0) are \mathcal{S} -generic.

Then $b_1 \sim a_0 < b_0 \sim a_1 < b_1$, a contradiction.

CASE 2. There are $a \in A_0, b \in B_0, aEb$ and $a < b$.

Exactly as in Case 1.

CASE 3. There are $a \in A_0, b \in B_0, aEb$ and $a \not\prec b$.

Let $D = \{(a, b) : a \in A_0, b \in B_0, aEb \text{ and } a \not\prec b\}$. Then II wins $G^2(D, E, \not\prec)$, say by \mathcal{T} .

Let $(a_0, b_0), (a_1, b_1), (a_2, b_2)$ be such that each (a_i, b_i) is \mathcal{T} -generic and (a_i, b_j) are \mathcal{S} -generic for $i \neq j$.

Then $a_0 \sim b_1 \sim a_2 \sim b_0$. But $a_0 \not\prec b_0$, a contradiction.

Hence, for all $a \in A, b \in B$ aEb implies $a \sim b$. Let $A' = \{a \in A : \exists b \in B \ aEb\}$. Then for all $a, a' \in A'$, aEa' implies $a \sim a'$. But then by Lemma 5.1, for all $a \in A'$, and all b, aEb implies $a \sim b$. But this contradicts $A' \subseteq Z$.

By Claims I and II, for all $Y \subseteq Z$, II has a winning strategy in $G(Y, Y, E, \approx)$. But then, on a dense set of $Y \subseteq Z$, I wins $G(Y, Y, E, \approx)$. By Lemma 4.2, \leq has a perfect set of incomparable elements, a contradiction. Hence $Z = \emptyset$ and we have established the main claim and proved Theorem 1.2.

The author would like to thank Donald Martin for remarking that since the argument given here uses only the fact that Γ is closed under universal real quantification, we can replace Γ by a smaller class Γ' . We can take Γ' to be the smallest adequate class of subsets of $(\mathbb{R} \cup \lambda)^n$ containing $T_{\leq}, T_{\not\prec}$, a function $\langle \cdot, \cdot \rangle : \lambda \times \lambda \rightarrow \lambda$ and closed under substituting parameters in λ and under universal real quantification. Let κ' be the supremum of the lengths of pre-wellorderings of \mathbb{R} in $\Gamma' \cap \check{\Gamma}'$. Then the proof given above shows that there is an $F : \mathbb{R} \rightarrow 2^\alpha$, some $\alpha < \kappa'$, such that if $x \leq y$ and $y \not\prec x$ then $F(x) <_{\text{lex}} F(y)$.

6. The Dilworth theorem

We now show that if \leq is thin, then $\mathbb{R} = \bigcup_{\alpha \in \gamma} Y_\alpha$, $\gamma \leq \lambda$ where

- (a) Y_α is pre-linearly ordered by \leq
- (b) $\{(\alpha, x) : x \in Y_\alpha\} \in \Delta$.

The latter statement implies that each Y_α is in Δ and hence is Suslin. Theorem 1.1 follows.

Let $\mathcal{O} = \bigcup \{Y \in \check{\Gamma} : \leq \text{pre-linearly orders } Y\}$. By Corollary 2.5, $\mathcal{O} = \bigcup \{Y \in \Delta : \leq \text{prelinearly orders } Y\}$. Hence, $\mathcal{O} \in \Gamma$. If $\mathcal{O} = \mathbb{R}$, we are done. Otherwise, let $W = \sim \mathcal{O}$. Then for all $Y \in \check{\Gamma} \cap \mathcal{P}(W)$, Y is not pre-linearly ordered by \leq .

Let $Y \in \check{\Gamma} \cap \mathcal{P}(W)$. Suppose that for all $x, y \in Y$, $x E_Y y$ implies $x \sim y$. Then for all $x, y \in Y$, if $x \not\sim y$ there is an $f \in \mathcal{F}_S^Y$ such that $f(x) < f(y)$ or $f(y) < f(x)$. Hence, $x \leq y$ or $y \leq x$ and thus Y is linearly ordered.

Hence, for all $Y \subseteq W$, $Y \in \check{\Gamma}$ there are $x, y \in Y$, $x E_Y y$ and $x \not\sim y$.

Since \leq is thin there is a dense open set in $\mathcal{P}(W) \cap \check{\Gamma}$ of non-empty Y such that II wins $G(Y, Y, E_Y, \cong)$ (Lemma 4.2). We work with Y in this dense open set.

Fix such a Y . Let $Y' = \{a \in Y : \exists b \in Y (a E_Y b \text{ and } a \not\sim b)\}$. Then $Y' \neq \emptyset$ and $Y' \in \check{\Gamma}$.

CLAIM 1. II has a winning strategy in $G(Y', Y', E_Y, \not\sim)$.

PROOF. Otherwise, let \mathcal{S} be I's winning strategy and (A, B) be I's first move.

CASE 1. There are $a \in A, b \in B$ $a E_Y b$ and $a \not\sim b$.

Let $D = \{(a, b) : a \in A, b \in B, a E_Y b \text{ and } a \not\sim b\}$. Then, since \cong is λ -Suslin, II has a winning strategy \mathcal{T} in $G^2(D, E_Y, \not\sim)$.

Choose $(a_0, b_0), (a_1, b_1)$ and (a_2, b_2) such that each (a_i, b_i) is \mathcal{T} -generic and if $i \neq i'$ then (a_i, b_i) is \mathcal{S} -generic. Then $a_0 \sim b_1 \sim a_2 \sim b_0 \not\sim a_0$, a contradiction.

CASE 2. Otherwise.

Since (A, B) is a legal move in $G(Y', Y', E_Y, \not\sim)$, $A' = \{a \in A : \exists b \in B, a E_Y b\} \neq \emptyset$. But if $a_1, a_2 \in A'$ and $a_1 E_Y a_2$ then there is a $b \in B$,

$a_1 E_Y b E_Y a_2$. So $a_1 \sim b \sim a_2$. Hence, for $a_1, a_2 \in A'$, $a_1 E_Y a_2$ implies $a_1 \sim a_2$. Since $E_{A'}$ refines E_Y , this implies A' is pre-linearly ordered by \leq , contradicting the definitions of W .

We now meld some strategies:

CLAIM 2. Let $A, B \subseteq Y'$, $A, B \in \tilde{I}$, with $(A \times B) \cap E_Y \neq \emptyset$. Then either I has a winning strategy in $G(A, B, E_Y, \not\prec)$ or I has a winning strategy in $G(A, B, E_Y, \succ)$.

PROOF. Suppose II wins $G(A, B, E_Y, \not\prec)$, with strategy \mathcal{S}_1 . Since II wins $G(Y', Y', E_Y, \approx)$, II has a winning strategy \mathcal{S}_2 in $G(A, B, E_Y, \geq)$. Since II wins $G(Y', Y', E_Y, \prec)$, II has a winning strategy \mathcal{S}_3 in $G(A, B, E_Y, \prec)$.

Melding these strategies we see that II has a winning strategy in $G(A, B, E_Y, <)$. Hence, by using II's strategy I wins $G(A, B, E_Y, \succ)$. \square

CLAIM 3. Let $A, B \in \mathcal{P}(Y) \cap \tilde{I}$ be such that $(A \times B) \cap E \neq \emptyset$. Let \mathcal{S}_1 be a winning strategy for I in $G(A, B, E_Y, \not\prec)$ (resp. $G(A, B, E_Y, \succ)$). Let $(A_0, B_0) = \mathcal{S}_1(\emptyset)$. Then for all $a \in A_0, b \in B_0$, if $a E_Y b$ then $a < b$ (resp. $a > b$).

PROOF. Otherwise, let $A^* = \{a \in A_0: \text{there is a } b \in B_0 \text{ } a E b \text{ and } a \not\prec b\}$. By Claim 2, I either wins $G(A^*, A^*, E_Y, \not\prec)$ or I wins $G(A^*, A^*, E_Y, \succ)$ (in fact, by symmetry, I wins both games). Without loss of generality let \mathcal{S}_2 be a winning strategy for I in $G(A^*, A^*, E_Y, \succ)$.

Let $\mathcal{S}_2(\emptyset) = (A', B')$. Let $D = \{(a, b): a \in B', b \in B_0, a E b \text{ and } a \not\prec b\}$. Since $B' \subseteq A^*$, $D \neq \emptyset$. Let \mathcal{T} be a winning strategy for II in $G^2(D, E_Y, \not\prec)$ (\mathcal{T} exists since $\not\prec$ is Suslin).

By Lemma 4.4, there are reals a_0, a_1, b such that (a_0, b) is \mathcal{S}_1 -generic (so $a_0 < b$) and (a_0, a_1) is \mathcal{S}_2 -generic (so $a_0 > a_1$) and (a_1, b) is \mathcal{T} -generic (so $a_1 \not\prec b$). But then $b > a_0 > a_1$ but $a_1 \not\prec b$ a contradiction. \square

By Claim 2, I either has a winning strategy in $G(Y', Y', E_Y, \not\prec)$ or $G(Y', Y', E_Y, \succ)$. Assume that I has winning strategy \mathcal{S}^* in $G(Y', Y', E_Y, \not\prec)$. Let $(A, B) = \mathcal{S}^*(\emptyset)$. By Claim 3 (A, B) satisfies the hypothesis of Lemma 3.1 (\leq -separation). Hence, we can find a $C \in \Delta$, $A \subseteq C$ and $C \cap B = \emptyset$ and C is downward closed in each equivalence class it intersects.

CLAIM 4. For all $x \in C \cap Y'$ and all $y \in \tilde{C} \cap Y'$, $x E_Y y$ implies $x < y$.

PROOF. Let $D = \{(d, c) : d \in \tilde{C} \cap Y', c \in C \cap Y', cE_Y d \text{ and } c \not\prec d\}$. Since $\not\prec$ is Suslin, II wins $G^2(D, E_Y, \not\prec)$ by a strategy \mathcal{T} .

Let $D^* = \pi_0(D)$, $C^* = \pi_1(D)$. By Claim 2, I either wins $G(D^*, D^*, E_Y, \not\prec)$ or $G(D^*, D^*, E_Y, \prec)$. By symmetry, I wins both. Let \mathcal{S}_1 be I's winning strategy in $G(D^*, D^*, E_Y, \not\prec)$. Let $(D_0, D_1) = \mathcal{S}_1(\emptyset)$.

Let $C^{**} = \{c \in C^* : \exists d \in D_1(c, d) \in D\}$.

Consider $G(D_0, C^{**}, E_Y, \not\prec)$ and $G(D_0, C^{**}, E_Y, \prec)$. By Claim 2, I wins one of these games.

Suppose I wins $G(D_0, C^{**}, E_Y, \prec)$. Let (D', C') be I's first move and let $d \in D', c \in C', dE_Y c$. By Claim 3, $d < c$. But C is downward closed in each E_Y equivalence class, a contradiction.

Hence, I wins $G(D_0, C^{**}, E_Y, \not\prec)$, by a strategy \mathcal{S}_2 .

By arguments similar to Lemmas 4.4, 4.3, we can find reals d_0, d_1, c such that (d_0, d_1) is \mathcal{S}_1 -generic, (d_0, c) is \mathcal{S}_2 -generic and (d_1, c) is \mathcal{T} -generic.

Hence, $d_0 < d_1, d_0 > c$ and $d_1 \not\prec c$. But then $c < d_0 < d_1$ so $c < d_1$, a contradiction. This establishes Claim 4. \square

By the glue together Lemma 3.3 we get that for all $a \in A, b \in B, a \not\prec_Y b$. Hence, (A, B) is not a legal move in $G(Y', Y', E_Y, \not\prec)$. But \mathcal{S}^* was supposed to be a winning strategy, a contradiction. This establishes the "Dilworth Theorem". \square

7. Miscellaneous remarks

We first remark that if \leq is thin and an equivalence relation, (i.e. $x \leq y$ iff $y \leq x$) and $\not\prec$ is λ -Suslin then the proof of §6 implies that \leq has $\leq \lambda$ many classes.

Namely, let W be as defined in the beginning of §6. By 4.2, there is a dense set of $Y \subseteq W$ such that II wins $G(Y, Y, E_Y, \sim)$, say by \mathcal{S} . Let $D = \{(a, b) : a, b \in Y \text{ and } a \not\prec b\}$. Since $\not\prec$ is λ -Suslin II wins $G^2(D, E_Y, \not\prec)$ say by \mathcal{T} .

Let $\langle a_{ij} : i \in 3, j \in 2 \rangle$ be such that (a_{i0}, a_{i1}) is \mathcal{T} -generic and for $i \neq i'$ $(a_{i0}, a_{i'1})$ are \mathcal{S} -generic. Then $a_{00} \sim a_{11} \sim a_{20} \sim a_{01}$ but $a_{00} \not\prec a_{01}$ a contradiction. Hence, $W = \emptyset$.

This fact is known as **Martin's Conjecture**. The arguments in HARRINGTON and SAMI (1978) also prove this modulo the results of Moschovakis.

In HARRINGTON *et al.* (unpublished) there is an argument that shows, in

particular, that there is no Borel Suslin line. It doesn't generalize directly under A.D. leading to the following question: Does $L(\mathbb{R}) \models$ "there is a Suslin line"?

Under large cardinals, this question is forcing absolute (see FOREMAN *et al.* (1988)). The author conjectures that the answer is "no."

References

- FOREMAN, M. and MAGIDOR, M., *The effect of large cardinals on the value of θ* (In preparation).
- FOREMAN, M., MAGIDOR, M. and SHELAH, S., *Martin's maximum, saturated ideals and non-regular ultrafilters*, Annals of Maths. 127, pp. 1–47.
- HARRINGTON, L. and KECHRIS, A., 1972, *On the determinacy of games on ordinals*, Annals of Math. Logic 4, pp. 229–308.
- HARRINGTON, L., MARKER, D. and SHELAH, S., *Borel orderings* (Unpublished manuscript).
- HARRINGTON, L. and SAMI, R., *Equivalence relations, projective and beyond*, in: Logic Colloquium '78' (North-Holland) pp. 247–264.
- HARRINGTON, L. and SHELAH, S., 1982, *Counting equivalence classes for co- κ -Suslin relations*, in: Logic Colloquium '80' (North-Holland).
- MARTIN, D., 1983, *The real game quantifier propagates scales*, Cabal Seminar 79–81, Lecture Notes in Math. 1019 (Springer-Verlag) pp. 157–171.
- MOSCHOVAKIS, Y., 1980, *Descriptive Set Theory* (North-Holland).

5 General Logic

This Page Intentionally Left Blank

LOGIC AND PRAGMATIC TRUTH

NEWTON C.A. DA COSTA

Department of Philosophy, University of Sao Paulo, Sao Paulo, Brazil

Introduction

This is basically an expository paper, in which I report some aspects of the work done by I. Mikenberg, R. Chuaqui, S. French and myself in the field of pragmatic truth. My exposition is based on MIKENBERG *et al.* (1986), DA COSTA and CHUAQUI (1989), DA COSTA (1986), DA COSTA (1987), and DA COSTA and FRENCH (1989a,b). I intend to show that the concept of pragmatic truth, at least in one of its possible interpretations, can be treated mathematically; the outcome is a formalization of that concept, analogous, in spirit, to Tarski's version of the classical, correspondence theory of truth.

There are three main conceptions of pragmatic truth, to wit: those of Peirce, of James, and of Dewey. Though James' and Dewey's conceptions are interesting and deserve to be studied in more detail, I will here be concerned only with Peirce's stance. (Dewey's notion of *warranted assertibility* constitutes a good candidate for a mathematical analysis, involving even the techniques of chronological logic.) Notwithstanding this, I do not want to make an exegesis of Peirce's work. On the contrary, his principal ideas will be, for me, only a motivation to develop some new views on the concept of pragmatic truth. Maybe the expert will consider that my formulation of this concept does really not capture the characteristic traits of Peirce's conceptions of truth. However, the definition of pragmatic truth contained in this article seems to me to be quite important, and its importance derives from its intrinsic merits, independently of exegetical questions.

Peirce wrote that, ". . . there is no distinction of meaning so fine as to consist in anything but a possible difference of practice." ["How to make our ideas clear", 1878, reproduced in HARTSHORNE *et al.* (1931–1958).]

He also declared that, "Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object." (*Ibidem.*) Another presentation of the same idea is this: "In order to ascertain the meaning of an intellectual conception one should consider what practical consequences might conceivably result by necessity from the truth of that conception; and the sum of these consequences will constitute the entire meaning of the conception." [HARTSHORNE *et al.* (1931–1958), Vol. V, paragraph 9.]

When Peirce talks of pragmatic truth, he is making reference to the *results* of scientific inquiry. According to him, "The opinion which is fated to be ultimately agreed to by all who investigate, is what we mean by truth . . ." ("How to make our ideas clear"), and if opinion "were to tend indefinitely toward absolute fixity", we would arrive at truth ["What pragmatism is", 1905, HARTSHORNE *et al.* (1931–1958).] Inquiry is controlled by the scientific community, being a social task. Therefore, it seems reasonable to suppose that *practice* can be identified with a collection of *primary statements*, which one can use to test, between certain limits, the propositions (theories and hypotheses) obtained in the way of inquiry. Sometimes, the class of primary statements may include general sentences, such as theories already incorporated into the body of science. A hypothesis is pragmatically true when it does not have consequences that contradict a primary statement; in addition, the better the hypothesis, the more it predicts. Put another way, a proposition is pragmatically true when things happen as if it were true (true in the sense of the correspondence theory of truth).

It also seems reasonable to accept that for some contingent propositions, which we will call *basic* or *decidable*, truth and pragmatic truth do coincide. In addition, a basic statement must be such that its truth or falsehood can, at least in principle, be settled. (Examples of basic propositions: "There is a pink pen on that table" and "Horse number two will win the Derby".) The classical concept of truth, then, constitutes one of the foundations of pragmatic truth in accordance with my conception (but other interpretations of the technical definition to be formulated are possible). For obvious reasons, the true decidable statements are always supposed to be included in the class of primary statements.

There are numerous situations, in the field of the empirical sciences, in which the concept of pragmatic truth can find applications. Three among them are the following: (1) Classical mechanics is at present known to be false. It was surmounted by relativistic mechanics. However, it can be

applied in several domains, within appropriate limits. This occurs, for instance, in engineering, where nobody would suggest the use of relativity. For the engineer, then, everything happens as though classical mechanics were true, i.e. as if it pictured reality. In other words, it is, for the engineer, pragmatically true. (2) We may conclude, as a lesson of the history of science, that experience, in the wide acceptance of the word, will sooner or later refute any theory as an absolutely true picture of reality. Owing to this circumstance, as well as to many others, it is best not to envisage a theory as true, but as pragmatically true. In the appraisal of hypotheses and theories, for instance, such a position is more reasonable than to take them as literally true [cf. DA COSTA 1986, DA COSTA 1987]. (3) Sometimes we conceive theories simply as instruments to save the appearances or calculating devices in relation to observation sentences. In this case, we are actually saying that theories can be, at most, pragmatically true. So, in order to articulate and systematize this stance, it seems interesting to have a developed, previously elaborated theory of pragmatic truth to begin with.

As I noted above, there may be authors who will think that the term 'pragmatic truth' is not being employed here in the historically correct way. Perhaps it would be more convenient if I were to use "quasi-truth" instead of "pragmatic truth" ("quasi-true" instead of "pragmatically true", etc.). Though this point constitutes only a minor terminological question for the objective of this paper, to emphasize its non-exegetical nature I will employ terms like "pragmatic truth" and "quasi-truth" as synonymous (see DA COSTA 1986).

1. Pragmatic structures and pragmatic truth

A statement is quasi-true or quasi-false (i.e., not quasi-true) only in relation to a given domain of knowledge, and within fixed limits of applicability. This fact is a corollary to the preceding informal discussion. Similarly, according to Tarski's characterization of the classical concept of truth, a proposition is true (or false) only in connection with an interpretation of its language in an appropriate, semantic structure. Precisely as the common concept of a structure underlies Tarski's definition of truth, the notion of simple pragmatic structure (sps, to abbreviate) constitutes the fundamental brick of the definition of quasi-truth or of pragmatic truth.

Let us suppose that we are studying a given domain of knowledge Δ . If we are trying to save the appearances in the domain, we should not be

able to derive, from our theories and hypotheses, statements which contradict any primary statement, whose set I shall denote by \mathcal{P} . Besides, the objects of Δ may be collected in a set \mathcal{A} , called the universe of the domain. The set \mathcal{A} may contain not only the actual objects of Δ , but also some extra, ideal objects which we postulate, to cope with Δ in an easier way. So, the universe \mathcal{A} encompasses real objects and ideal ones (in some cases, merely fictitious objects), that are linked by certain relations, in particular by monadic relations. These relations constitute the *atomic* inter-connections bearing among the elements of \mathcal{A} . I conceive these relations as the semantic counterpart of what we accept as true or what we know as true about the atomic links existing among the members of the universe. In consequence, such relations must in general be partial relations. An n -ary partial relation, R , $0 < n < \omega$, is an ordered pair $\langle R_1, R_2 \rangle$ satisfying the conditions that, (1) if an n -tuple of elements of \mathcal{A} belongs to R_1 , then it is said also to belong to R and its terms to bear the relation R , and (2) if an n -tuple belongs to R_2 , then it is said not to belong to R and its terms not to be linked by R . When an n -tuple belongs neither to R_1 nor to R_2 , we say that R is not defined for that n -tuple. In other words, an n -ary partial relation is not necessarily *defined* for all n -tuples of elements of the universe. Obviously, a common, total relation is a particular case of partial relation.

As a result of our discussion, I present the following definition, that formally characterizes the concept of a sps:

DEFINITION 1. A simple pragmatic structure (sps) is a triple

$$\mathfrak{U} = \langle \mathcal{A}, R_i, \mathcal{P} \rangle_{i \in I}$$

where \mathcal{A} is a non-empty set, called the universe of \mathfrak{U} , R_i is a partial relation on \mathcal{A} for every $i \in I$, and \mathcal{P} is a set of sentences (closed formulas) of a language \mathcal{L} of the same similarity type as that of \mathfrak{U} , and which is interpreted, in an obvious sense, in \mathfrak{U} . (The case in which R_i is empty, for some $i \in I$, or in which $\mathcal{P} = \phi$ are not excluded.)

We now proceed to formulate the definition of pragmatic truth (or of quasi-truth), the central notion of this paper. But before we need a preliminary concept.

DEFINITION 2. $\mathfrak{U} = \langle \mathcal{A}, R_i, \mathcal{P} \rangle_{i \in I}$ and \mathcal{L} will denote a sps and a first-order language respectively, the latter interpreted in the former. \mathfrak{B} will

designate a usual, total structure (whose n -adic relations are standard relations, defined for all n -tuples of the universe), in which \mathcal{L} is also interpreted. \mathfrak{B} is called \mathfrak{A} -normal if the following conditions are satisfied:

- (1) The universe of \mathfrak{B} is \mathcal{A} ;
- (2) The (total) relations of \mathfrak{B} extend the corresponding partial relations of \mathfrak{A} ;
- (3) If c is an individual constant of \mathfrak{B} , then in both \mathfrak{A} and \mathfrak{B} c is interpreted by the same element;
- (4) If $\alpha \in \mathcal{P}$, then $\mathfrak{B} \models \alpha$.

To simplify the exposition, I suppose that \mathcal{L} does not contain function symbols. It is clear that the set of \mathfrak{A} -normal structures may be empty. A necessary and sufficient condition for the non-emptiness of this set is found in MIKENBERG *et al.* (1986). I shall presuppose that all sps's considered here satisfy that condition.

DEFINITION 3. Let \mathfrak{A} and \mathcal{L} be as in the preceding definition, and α a sentence of \mathcal{L} . α is said to be pragmatically true (quasi-true) in \mathfrak{A} according to \mathfrak{B} if \mathfrak{B} is an \mathfrak{A} -normal structure and α is true in \mathfrak{B} (in the Tarskian sense); α is said to be pragmatically true (quasi-true) in the sps \mathfrak{A} if there exists an \mathfrak{A} -normal structure \mathfrak{B} in which α is true. If α is not pragmatically true or not quasi-true in the sps \mathfrak{A} according to \mathfrak{B} (is not pragmatically true or quasi-true in the sps \mathfrak{A}), we say that α is pragmatically false or quasi-false in the sps \mathfrak{A} according to \mathfrak{B} (is pragmatically false or quasi-false in the sps \mathfrak{A}).

When we assert that a certain sentence (hypothesis, theory) α saves the appearances in a domain Δ or that things occur in Δ as if α were true, this can be formalized as follows. First of all, we replace Δ by a sps $\mathfrak{A} = \langle \mathcal{A}, R_i, \mathcal{P} \rangle_{i \in I}$, where \mathcal{A} is the universe of Δ , R_i , $i \in I$, are the partial relations which interest us, and \mathcal{P} is the set of primary statements. I insist in the fact that \mathcal{P} contains all propositions (of the language of \mathfrak{A}) which are known to be true or accepted as true, what normally includes laws and theories in force in Δ ; furthermore \mathcal{P} has to contain all decidable true statements, because they constitute an important part of the process of corroborating new hypotheses and theories in Δ . The interplay between Δ and \mathfrak{A} depends on some explicit or implicit rules which connect the elements of Δ with those of \mathfrak{A} . These rules involve questions of measurement and of statistical theory, and it seems that they are not always the same for all domains; for example, these rules apparently

differ in the cases of classical mechanics and of quantum mechanics. Sometimes, we have recourse to paradigmatical examples, that show how the possible applications should be accomplished and judged. However, we know how the concept of pragmatic truth can contribute to clarify some issues of the philosophy of science: a hypothesis saves the appearances in Δ if it is quasi-true in \mathfrak{A} or, loosely speaking, in Δ . Among the possible, pragmatically true hypotheses, we search for the most fertile, the most intuitive, etc. Concerning the choice of the best hypothesis, one may appeal to probability and to other devices (see DA COSTA 1986, DA COSTA 1987).

The applications of the physical theories, for example, always follow the above pattern, at least in principle, even when the theories are envisaged as true, and not as simply pragmatically true. As Dorling writes, in a different context, "What is important is not how philosophers construe physicists' theories but how physicists construe them. But it is philosophers, not physicists, who take a theory to be asserted as true for arbitrarily small space-time regions, throughout all space and time, for arbitrary extremes of conditions whose variation has not been studied experimentally, and so on, unless anything is explicitly said to the contrary. Physicists would normally assume that no such commitment is included unless it be explicitly asserted. So is there really a difficulty of the kind envisaged, when one comes to consider actual examples of theoretical issues on which the progress of physics has depended and of the sort which it would be instructive to subject to Bayesian analysis? Whether or not Venus's orbit encircles the sun? Whether or not the earth has an annual motion relative to the fixed stars? Whether or not the smaller parts of a gas apparently at rest are in reality in violent motion? Whether contiguous portions of palpably homogeneous bodies cease to be similar when their dimensions are of the order of a thousand-millionth of a centimetre? Whether or not the time-interval between two events is independent of the path along which it is measured? Whether a man falling in a lift could readily detect that his motion was non-inertial by observing the trajectory of a light ray? Whether energy is conserved in the processes of emission and absorption of light? Whether or not energy is conserved in beta-decay? Whether the four-fermion interaction contains equal amounts of vector and vector coupling? There seems no *prima facie* reason why the rational degree of belief in such propositions or their negations should always be zero or extremely small. Of course if, in deference to L.S.E. philosophy, we graft into such proposition implicit universal quantifications with respect to additional unmentioned vari-

ables, then we can no doubt persuade ourselves that now the probabilities must be zero or extremely small. But surely the study of scientific inference is concerned with the theories that scientists discuss and especially with the parts that scientists sometimes come to believe in, not with theories invented by philosophers that no one could ever possibly believe in.” (DORLING 1972: pp. 183–184.)

Perhaps distorting a little Dorling’s words, I will interpret them as meaning that acceptable scientific propositions are in intention pragmatically true, though in special situations they may be true *tout court*; besides, that nothing does hinder the use of (subjective) probabilistic techniques to evaluate the quasi-truth of those propositions (cf. DA COSTA 1986, DA COSTA 1987).

2. The logic of pragmatic truth

Given a sps \mathfrak{A} , the set of all normal \mathfrak{A} -models presents an obvious analogy to the worlds of a Kripke model. This remark leads us to extend the first-order language in which we talk about a sps by the adjunction of the modal operators \Box (necessity) and \Diamond (possibility). Let us call the resulting language \mathcal{L}' . Suppose that \mathfrak{A} is a fixed sps; in \mathcal{L}' we will take $\Box\alpha$ and $\Diamond\beta$ to represent, respectively, the statements that α is true in every \mathfrak{A} -normal structure (or, to abbreviate, true in \mathfrak{A}) and that β is true in some \mathfrak{A} -normal structure.

Let α be a sentence of \mathcal{L}' ; then it is clear that α is pragmatically valid, i.e. pragmatically true in every sps, if it is a theorem of S5 with quantification and necessary equality, a system which will be denoted by S5Q (see HUGHES and CRESSWELL 1968).

Given a formula α of \mathcal{L}' , $\forall\alpha$ will denote α preceded by an arbitrary sequence of quantifications, subject only to the restriction that in $\forall\alpha$ no variable is free.

In general, we define formally pragmatic validity as follows. A formula α of \mathcal{L}' is pragmatically valid if $\Diamond\forall\alpha$ is a theorem of S5Q. Therefore, α is pragmatically valid (or quasi-valid) when $\Diamond\forall\alpha$ is valid in the standard sense in S5Q. The motivation of this definition is quite evident.

I will designate by PT the logical system composed of the pragmatically valid formulas of \mathcal{L}' . Taking into account its definition, PT constitutes Jaśkowski’s discussive logic associated with S5Q (see JAŚKOWSKI 1969, DA COSTA 1975, DA COSTA and DUBIKAJTIS 1977).

The primitive symbols of \mathcal{L}' , the language of PT, are the following: (1) Connectives: \vee (or), \neg (not), and \Box (necessity); \rightarrow (implication), \wedge (conjunction), \leftrightarrow (equivalence), and \Diamond (possibility) are defined as usual. (2) The universal quantifier, \forall ; the existential quantifier, \exists , is introduced by the common definition. (3) Individual variables: a denumerably infinite collection of individual variables. (4) Constants: an arbitrary family of individual constants. (5) For any natural number n , greater than zero, an arbitrary family of predicate symbols of rank n . (6) The equality symbol: $=$. (7) Auxiliary symbols: parentheses.

We introduce the syntactic concepts, such as those of formula, term, variable free in a formula, etc., in the standard way.

Postulates (primitive rules and axiom schemes) of PT:

(I) If α is an instance of a (propositional) tautology, then α is an axiom

$$(II) \frac{\Box = \forall \alpha \Box \forall (\alpha \rightarrow \beta)}{\Box \forall \beta} \quad (III) \Box \forall (\Box(\alpha \rightarrow \beta) \rightarrow (\Box \alpha \rightarrow \Box \beta))$$

$$(IV) \Box \forall (\Box \alpha \rightarrow \alpha) \quad (V) \Box \forall (\Diamond \alpha \rightarrow \Box \Diamond \alpha)$$

$$(VI) \Box \forall (\forall x \alpha(x) \rightarrow \alpha(t)) \quad (VII) \frac{\Box \forall \alpha}{\alpha} \quad (VIII) \frac{\Box \forall \alpha}{\Box \forall \Box \alpha}$$

$$(IX) \frac{\Diamond \forall \alpha}{\alpha} \quad (X) \frac{\Box \forall (\alpha \rightarrow \beta(x))}{\Box \forall (\alpha \rightarrow \forall x \beta(x))}$$

(XI) Vacuous quantifications may be introduced or suppressed in any formula

$$(XII) \Box \forall x(x = x) \quad (XIII) \Box \forall (x = y \rightarrow (\alpha(x) \leftrightarrow \alpha(y)))$$

The preceding postulates are subject to the common restrictions. Furthermore, in PT the definitions of proof and of (formal) theorem are the usual ones.

Loosely speaking, in PT \Box and \Diamond may be interpreted as the operators of pragmatic validity and of pragmatic truth respectively.

We can prove that α is a theorem of PT if, and only if, $\Diamond \forall \alpha$ is a theorem of S5Q. This soundness and completeness result gives immediately a semantics for PT in terms of Kripke's models, relative to which it is sound and complete (for details, see DA COSTA and CHUAQUI 1989).

We verify, therefore, that PT, the logic of pragmatic truth or of

quasi-truth, extends classical first-order logic. In fact, the latter constitutes the basis of the former.

To finish this section, I observe that PT can be extended to a higher-order logic.

3. Philosophical considerations

We can immediately distinguish three inter-related areas in the philosophy of science to which the formalism of pragmatic truth can be applied: the realist-empiricist debate, the probabilistic approach to confirmation and the problem of induction, and, finally, the nature and structure of scientific theories in general.

As regards the first, it seems that the above theory is sufficiently malleable to serve either side in the discussion. Thus, taking only those members of \mathcal{A} which represent the actual, *observable* objects of Δ and also the sub-set of the set of primary statements which may be regarded as *observation* statements, we obtain a sub-structure which can be taken to represent the “empirical sub-structures” of van Fraassen’s “constructive empiricism” (VAN FRAASSEN 1980, FRENCH 1989a). These are then regarded as *embedded* in structures of the form \mathfrak{A} , the extra unobservable elements of which (contained in \mathcal{A}) are considered to be simply “convenient fictions” having a pragmatic value only, in the context of explanation and prediction. Since α being pragmatically true implies everything happening in Δ as if α were true, then α can be taken to “save the phenomena” in exactly the sense that constructive empiricism wants.

The difference between this and some form of realism turns, of course, on the attitude towards the unobservable elements of \mathcal{A} and the question as to how this attitude is licensed (SALMON 1984, Ch. 8). Regarding such elements as not merely ideal, in some sense, then opens up the possibility of using the above formalism as a platform on which to construct a form of “pragmatic realism” (FRENCH 1989b).

Any such attempt must take note of the distinction between the “metaphysical” and “epistemological” aspects of the realist programme (PUTNAM 1978). The first relates to the *nature of truth*, whereas the latter can be said to be concerned with *what is true*. Following the various well known criticisms of the correspondence theory of truth in this context, realists have retreated to the claim that mature theories (whatever that may mean) are typically “approximately” true (BOYD 1976, PUTNAM 1978). This position has in turn fallen under attack, principally because of

the perceived difficulties in formalising what it means for a theory to be “approximately” true. It can immediately be seen that the above formalism can be usefully employed in the defence of realism on this point, although its use impels us to recast realism into a different form, one that is both more sophisticated and closer to some middle way between its naive predecessor and pure instrumentalism.

Thus considering so-called “epistemological realism”, the important questions from the point of view outlined here are “what are to be included as members of \mathcal{A} ?” and “what is the nature of the domain Δ modelled by our simple pragmatic structure \mathcal{X} ?”. These are of course intimately linked.

Taking the second question first (cf. SHAPER 1977) I begin by emphasising the distinction between the domain of knowledge that a theory *should* account for and the domain that it *actually does* account for successfully (NICKLES 1977: p. 583). Relative to the latter a theory is pragmatically true *forever*: this is why we can still use classical mechanics today for building bridges etc. However it is through the first kind of domain that a theory’s deficiencies are illuminated and thus it is via the mismatch between these two types of “domain” that a theory is shown to be inadequate, leading to the search for a better one. The difference is therefore important because it effectively drives the machinery of theory change and scientific progress in general.

As regards the exact nature of these domains, and of the former kind in particular, I note that the set of distinguished sentences, or primary statements, P , will in general include both true decidable sentences such as observation sentences and certain general propositions encompassing laws and theories already assumed to be true. This already suggests that Δ is not merely a collection of phenomena but is *ordered* in some way by these latter laws, theories, symmetry principles etc. If science is taken to be cumulative, in the sense that each successive theory does not start afresh, as it were, but builds upon at least part of the structure(s) of its predecessors, then Δ may be taken to include these latter aspects as well (cf. SHAPER 1977). In this case the concept of a domain becomes a “pre-theoretical notion” in the sense that it is described, at least in part, in the pre-theoretical language of previous theories. Modelling Δ by \mathcal{X} may then actually give rise to a *different* domain to be modelled by subsequent theories. This is already a big concession to the realist point of view since if the successor theory entails that the previous one was strictly incorrect then for an instrumentalist there can be no reference to such a domain described in “pre-theoretical” terms (VAN FRAASSEN 1977: p. 589).

This brings us on to the question of the members of A and our attitude towards them. As I have said, \mathcal{A} will in general include both (directly) observable and non-observable (“theoretical”) terms. The problem of providing some warrant for the realist’s claim that some of the latter, at least, *refer*, is central to recent discussions of this issue. Realist arguments to the effect that such warrant is given by the success of theoretical predictions, indeed by the success of science itself, have been met by empiricist constructions of that success in terms of empirical adequacy, with successful prediction assigned a pragmatic role only (VAN FRAASSEN 1980, 1985).

On this point the empiricist may be usefully compared to the sceptic in his/her search for some “bed-rock” of knowledge, this being directly observable phenomena. One can argue, however, that knowledge is not gained by merely passively observing phenomena but by interacting with them (LUNTLEY 1982, HACKING 1983) and it is through such interaction with, and manipulation of, the phenomena that we may be said to have knowledge of so-called “unobservable” entities. It is difficult to see how such interacting and manipulating can be regarded as in any way rational unless there is a belief in the actual existence of such entities (cf. HORWICH 1987).

Of course, not all of the theoretical elements of \mathcal{A} refer to unobservable entities in this sense, since some will be “idealization terms”, such as “point particle” and “rigid rod” (SUPPE 1977: p. 568). Furthermore, those assertions which involve the “existence terms” of \mathcal{A} make claims of varying degrees of strength, depending on, among other things, the status of the theories in which they occur. The difference in this degree of strength corresponds to the difference in scientists’ degree of belief in such claims. Thus they include claims about entities: (i) that were once asserted to exist but are not now regarded as existing, such as phlogiston; (ii) whose existence or non-existence has not yet been established, e.g., tachyons; and (iii) which are now considered to exist, such as electrons or quarks.

“Naive” realism erred in lumping all such claims together and thus came to grief over the history of science. Any improved version worthy of the title must therefore provide an answer to the question of how these different attitudes towards the members of the “scientific zoo” can be accommodated (FINE 1984). The most obvious way in which to do this is to introduce degrees of belief, which permit exactly the kind of gradation in our attitude towards theoretical entities which a more sophisticated realism requires. Thus there are various theoretical claims in which we have a high degree of confidence, such as that electrons exist, others in

which we presently have very little confidence, for example involving phlogiston, and others for which our degrees of belief cover the whole spectrum in between these two extremes (cf. HORWICH 1982: p. 136).

The introduction of degrees of belief in this way obviously cuts across the naive realist-simple instrumentalist divide and brings us on nicely to our second application of the pragmatic truth formalism. Given our comments above about the "approximate" truth of scientific theories and scientists' belief attitudes towards them, it obviously makes sense to treat the belief in a theoretical proposition not as belief in its truth *per se* but as belief in its pragmatic truth. Identifying the degree of belief in the proposition with its subjective probability then leads to the "pragmatic probability" interpretation of the probability calculus (DA COSTA 1986). Thus the degree of belief in the pragmatic truth of α is identified with the subjective pragmatic probability of α .

Recent Bayesian accounts of such notions as simplicity, the desire for varied evidence, the supportive power of surprising predictions etc. (HORWICH 1982) can be easily accommodated within this interpretation. In particular we can give a straight forward account of the confirmation of hypotheses in terms of confirmatory evidence, garnered from the domain Δ of the theory, increasing our degree of belief in the pragmatic truth of the hypothesis, the change in degree of belief/pragmatic probability on the evidence proceeding according to a suitable form of Bayes' Theorem. This forms the basis of a "logic" of induction which nicely resolves such standard difficulties as the assignment of non-zero *a priori* probabilities to universal hypotheses and the so-called Hacking problem involving non-Bayesian changes in our degrees of belief (DA COSTA 1987, DA COSTA and FRENCH 1989a). The keystone of this system is the effective reduction of all inductions to the general hypothetico-deductive method. In the latter we have that certain propositions (the premisses or pieces of evidence) make plausible, in the light of certain side conditions, a new proposition, the conclusion or hypothesis. When the conjunction of the premisses is logically implied, by the hypothesis and the side conditions, we have the particular instance of the strict hypothetico-deductive method.

The generalised hypothetico-deductive method may thus be envisaged as, in a certain sense, the basic form of inductive inference: any induction whatsoever can be viewed as an application of this method. The significance of this move is that it allows us to probabilize all inductions. Since the conclusion of an induction has a tentative status and is regarded as quasi-true only, we should obviously employ the pragmatic probability calculus in this procedure. The reliability of an induction depends on the

plausibility conferred by the set of premisses on the conclusion, given the side conditions. This plausibility can then be estimated, whether qualitatively or quantitatively, by means of pragmatic probabilities. The system thereby developed is local, instrumental and tentative, three characteristics which it shares with Shimony's "tempered personalist" view (SHIMONY 1970).

Finally I come to the question of the nature of structure of scientific theories themselves. The introduction of simple pragmatic structures can obviously and conveniently be located within the semantic, or model-theoretic, approach to scientific theories first introduced by SUPPES (1957, 1967, 1970) and BETH (1948, 1961). For a general introduction see SUPPE (1977: pp. 221–230). The core idea of this approach is to consider a scientific theory in terms of a description of its set of models regarded as the structures it makes available for modelling its domain (cf. VAN FRAASSEN 1980, pp. 41–69). However, given that theories are not, and should not be, considered as literally true, this programma should be adapted to admit *partial* structures, such as are expressed by the simple pragmatic structures above (DA COSTA and FRENCH 1989b).

We have seen that perhaps the principal lesson to be drawn from the realist-empiricist debate is that, as regards the relationship between theories and "the world", we should adopt some kind of "fallibilist" position, in terms of a theory of truth other than the simple correspondence view. It is therefore clearly both more plausible and more rational to model our domain of knowledge by structures of the form \mathfrak{A} above, rather than by "complete" or fully specified constructions.

The model-theoretical approach in general gains a certain plausibility from the fact that "iconic" models are extensively used in science itself. Representing such models by simple pragmatic structures also (DA COSTA and FRENCH 1989) helps us to classify and understand the way they are used, whether for heuristic purposes, to qualitatively "probe" a complicated theory, or to test a theory when there is a computation gap (REDHEAD 1980). Thus the relationship between a theory, represented in model-theoretic terms, and its iconic models can be easily and conveniently explored through such an approach.

Representing a theory in this way also allows us to accommodate the notion of theoretical unification in science. Thus the unification of two theories T and T' might be achieved by the identification of elements of their sets of individuals A and A' , such elements then being said to have "unifying power" (FRIEDMAN 1983). Of course, due regard must be paid to the question of which elements are identified in this way and to the

relationship between the domains Δ and Δ' of the two theories. For example, joint application of T and T' may lead to certain "crossbreed" theoretical objects being posited in order to model the (new) domain of the unified theory (NUGAYEV 1985). Subsequent elimination of such crossbreeds can then be achieved through the construction of a further embracing theory T'' which contains both T and T' as derivative sub-theories or by reducing T to T' . The latter will typically occur when the domain of T is completely contained within that of T' .

However ontological unity is not the only possible notion of unity to be found in science (REDHEAD 1984) and we conjecture that these others may also be conveniently dealt with through this approach.

Hopefully this short exposition has given some idea of the power and fruitfulness of the twin concepts of pragmatic truth and simple pragmatic structure as regards characterising and understanding some of the more important aspects of the scientific process. As well as the problem of theory unification, possible future applications include theory evolution and inter-theoretical relations, the relationship between pragmatic truth and natural laws and the modelling of quantum mechanics in terms of simple pragmatic structures. And this is only just the beginning!

References

- BETH, E., 1948, *Natuurphilosophie* (Noorduyun, Gorinchem).
- BETH, E., 1961, *Semantics of physical theories*, in: H. Freudenthal, ed., *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences* (Reidel, Dordrecht), pp. 48–51.
- BOYD, R., 1976, *Approximate truth and natural necessity*, *Journal of Philosophy* 73, pp. 633–635.
- DA COSTA, N.C.A., 1975, *Remarks on Jaśkowski's discussive logic*, *Reports on Mathematical Logic* 4, pp. 7–16.
- DA COSTA, N.C.A., 1986, *Pragmatic Probability*, *Erkenntnis* 25, pp. 141–162.
- DA COSTA, N.C.A., 1987, *Outlines of a System of Inductive Logic*, *Teoria* 7, pp. 3–13.
- DA COSTA, N.C.A. and DUBIKAJTIS, L., 1977, *On Jaśkowski's discussive logic*, in: A.I. Arruda, N.C.A. da Costa and R. Chuaqui, eds., *Non Classical Logics, Model Theory and Computability* (North-Holland), pp. 37–56.
- DA COSTA, N.C.A. and CHUAQUI, R., 1989, *The Logic of Pragmatic Truth* (to appear).
- DA COSTA, N.C.A. and FRENCH, S., 1989a, *Pragmatic Truth and the Logic of Induction* (to appear).
- DA COSTA, N.C.A. and FRENCH, S., 1989b, *The Model-Theoretic Approach in Philosophy of Science*, to appear in *Philosophy of Science*.
- DORLING, J., 1972, *Bayesianism and the rationality of scientific inference*, *British Journal for the Philosophy of Science* 23, pp. 181–190.
- FINE, A., 1984, *And not anti-realism either*, *Noûs* 18, pp. 51–56.
- FRENCH, S., 1989a, *A Note on Constructive Empiricism and Pragmatic Truth* (to appear).

- FRENCH, S., 1989b, *Pragmatic Realism* (to appear).
- FRIEDMAN, M., 1983, *Foundations of Space-Time Theories* (Princeton University Press).
- HACKING, I., 1983, *Representing and Intervening* (Cambridge University Press).
- HARTSHORNE, C., WEISS, P.C. and BURKS, A. eds., 1931–1958, *Collected Papers of C.S. Peirce*, 8 vols (Harvard University Press).
- HORWICH, P., 1982, *Probability and Evidence* (Cambridge University Press).
- HORWICH, P., 1987, *Does believing a theory take more than just using it?*, paper read at the VIII International Congress of Logic, Methodology and Philosophy of Science, Moscow 17–22 August, 1987.
- HUGHES, C.H. and CRESSWELL, M.J., 1968, *An Introduction to Modal Logic* (Methuen).
- JAŚKOWSKI, S., 1969, *Propositional calculi for contradictory deductive systems*, *Studia Logica* 26, pp. 143–157.
- LUNTLEY, M., 1982, *Verification, perception and theoretical entities*, *Philosophical Quarterly*, 32, pp. 245–261.
- MIKENBERG, I., DA COSTA, N.C.A. and CHUAQUI, R., 1986, *Pragmatic truth and approximation to truth*, *The Journal of Symbolic Logic* 51, pp. 201–221.
- NICKLES, T., 1977, *Heuristics and justification in scientific research*, in: F. Suppe, ed., *The Structure of Scientific Theories*, 2nd edn. (Univ. of Illinois Press), pp. 571–589.
- NUGAYEV, R., 1985, *A study of theory unification*, *British Journal for the Philosophy of Science* 36, pp. 159–173.
- PUTNAM, H., 1978, *Meaning and the Moral Sciences* (Routledge and Kegan Paul, London).
- REDHEAD, M., 1980, *Models in Physics*, *British Journal for the Philosophy of Science* 31, pp. 145–163.
- REDHEAD, M., 1984, *Unification in science* (review of C.F. von Weizsäcker, *The Unity of Nature*), *British Journal for the Philosophy of Science* 35, pp. 274–279.
- SALMON, W., 1984, *Scientific Explanation and the Causal Structure of the World* (Princeton University Press).
- SHAPER, D., 1977, *Scientific Theories and Their Domains*, in F. Suppe, *op. cit.*, pp. 518–565.
- SHIMONY, A., 1970, *Scientific inference*, in: R.G. Colodny, ed., *The Nature and Function of Scientific Theories* (Univ. of Pittsburgh Press), pp. 79–172.
- SUPPE, F., 1977, *Editorial interpolation: Shapere on the instrumentalistic vs. realistic conceptions of theories*, in F. Suppe, *op. cit.*, pp. 566–570.
- SUPPES, P., 1957, *Introduction to Logic* (van Nostrand, New York).
- SUPPES, P., 1967, *What is a Scientific Theory?* in: S. Margenbesser, ed., *Philosophy of Science Today* (Basic Books, New York), pp. 55–67.
- SUPPES, P., 1970, *Set Theoretical Structures in Science*. Mimeographed lecture notes (University of Stanford, Stanford).
- VAN FRAASSEN, B., 1977, *Discussion*, in: F. Suppe, ed., *The Structure of Scientific Theories*, *op. cit.* pp. 598–599.
- VAN FRAASSEN, B., 1980, *The Scientific Image* (Oxford University Press, Oxford).
- VAN FRAASSEN, B., 1985, *Empiricism in the Philosophy of Science*, in: P. M. Churchland and C.A. Hooker, eds., *Images of Science* (University of Chicago Press, Chicago), pp. 245–368.

This Page Intentionally Left Blank

THE JUSTIFICATION OF NEGATION AS FAILURE

KIT FINE

Department of Philosophy, UCLA, Los Angeles, CA 90024, USA

Prolog is a logic programming language; it is used to answer queries on the basis of information provided by the programmer. For the most part, the logic employed by Prolog is standard. But it uses a highly unorthodox rule for establishing negative facts. This rule, the so-called rule of *negation as failure*, allows us to deny a statement on the grounds that a certain attempt to prove it has failed.

The rule is not classically valid; and therefore the question arises as to how it is to be justified. There are basically three different kinds of justification that have been proposed in the literature. The first is to re-interpret negation to mean something like unprovability. The second is to assume that the program is complete with respect to truths; all truths are derivable. The third is to suppose that the program is complete with respect to conditions; all sufficient conditions for the application of the predicates have been specified.

My aim in this paper is to evaluate these various proposals and then to make a proposal of my own. I shall argue that the existing proposals all suffer from some defect or another: the first is unable to account for a classical reading of negation; the second delivers too much on programs which employ negation; and the third delivers too little on programs which make no use of negation.

I shall then argue that my own proposal is able to avoid these difficulties. From one point of view, the proposal is not new; it is merely a form of the second proposal stated above, according to which all truths are derivable. However, the concept of derivability which is appealed to is quite novel; for the assumption that all truths are derivable, may itself be used in establishing that a given statement is derivable. The assumption has, in other words, a self-referential character.

The proposal has various other features of interest. It provides a

natural way of interpreting inductive definitions in which the positive instances of a predicate are allowed to depend upon its negative instances. It sanctions an extension of the rule of negation of failure, under which not only the finite, but also the transfinite, failure of a statement may constitute a ground for its denial. It is capable of variation in the choice of which other assumptions or rules are used in defining the concept of derivability.

I have tried to make the paper accessible to the general reader. For this reason, I have included a description of Prolog, or rather of that idealized version of Prolog which is most relevant to our concerns. A general introduction to the foundations of logic programming is given by LLOYD (1984), and a survey of recent work on the justification of negation as failure is given by SHEPHERDSON (1988). I have for the same reason suppressed most of the technical details, including those concerning the complexity of my method and the cases of agreement with other methods. I hope to give a fuller technical account elsewhere.

One feature of my exposition is worthy of special note. I have for the most part confined my attention to the sentential case, under which only truth-functional complexity is ever exposed. Such a case is usually regarded as trivial, since most of the interesting features of Prolog depend upon the use of variables. However, in this regard, the rule of negation as failure is an exception. Most of the problems in justifying the rule already arise at the sentential level; and to solve these problems at this level is to have gone a long way towards solving them altogether. There are, however, certain difficulties which are peculiar to the introduction of variables and terms; and these are considered at the end of the paper. It is argued, in particular, that the usual assumptions concerning an ontology of terms are needlessly strong and that an ordinary ontology of individuals can be countenanced in its place.

1. What is Prolog?

Prolog may be viewed as a mechanism for answering queries. In response to a query of the form "Is such and such a statement true?", it returns the answer "Yes" or "No" (if it returns any answer at all); and in response to a query of the form "Which individuals satisfy such and such a condition", it provides specifications of the individuals (if it provides any specification at all).

Prolog answers these queries on the basis of the information embodied

in a *program*. A program is a set of clauses or claims. However, there are severe limitations on the language within which the clauses and queries of Prolog are to be formulated. One would *like* to be able to make any claim and to ask any question; but one also wants to compute the answers to the questions asked. The language of Prolog represents a compromise, and a happy one at that, between these competing demands of expressivity and efficiency.

The queries and the clauses must, in the first place, be constructed from atoms. An *atom* is the result of applying a predicate to an appropriate number of terms; and a term is either a variable or the result of applying a function symbol to an appropriate number of other terms. It is allowed that a predicate may apply to no terms (in which case it is simply a sentence-letter); and it is allowed that a function symbol may apply to no terms (in which case it is simply a constant).

A *literal* is either an atom or the negation of an atom. A *clause* is then a conditional to the effect that an atom holds if certain literals hold. It may be written in the form:

$$B \leftarrow A_1, \dots, A_n$$

$n \geq 0$, where B is the atom at the *head* of the conditional and A_1, \dots, A_n are the literals of its *body*. In orthodox logical notation, the clause would be written as:

$$(A_1 \& \dots \& A_n) \rightarrow B.$$

The clause is said to be *categorical* if $n = 0$ and *conditional* otherwise. A *query*, on the other hand, is either a request for the truth-value of a conjunction of literals, should they contain no variables, or a request for a specification of the individuals which satisfy the conjunction, should they contain variables. It may be written as:

$$?A_1, \dots, A_n$$

$n \geq 0$, where A_1, \dots, A_n are the component literals.

A literal, clause, program, or query is said to be *positive* if it contains no occurrences of the symbol “-” for negation; and otherwise it is said to be *negative*. A literal, clause, program, or query is said to be *ground* if it contains no variables; and otherwise it is said to be *unground*. In fact, many of the important features of Prolog do not depend on the internal structure of the atoms or literals; and for this reason, I shall sometimes talk of *statements* instead of ground atoms or of ground literals.

Prolog attempts to answer a query by working backwards. A query $?A_1, \dots, A_n$ is treated in effect as the goal: prove an answer to the query. The clauses of the program are then used successively to reduce this goal to further goals. One question leads to another until eventually (it is hoped) an answer is found.

Thus central to the operation of Prolog is a mechanism for reducing one goal to another. Let us for the moment assume that all of our programs and queries are positive (it is as if negation had been banished from the language). Then in the case in which no variables are present in the query or in the clauses of the program, the method of reduction takes an especially simple form. For the query $?A_1, \dots, A_n$ asks whether $A_1 \& \dots \& A_n$ is true, and the corresponding goal is to prove $A_1 \& \dots \& A_n$, i.e. to prove each of A_1, \dots, A_n . Therefore, given that $A_i \leftarrow B_1, \dots, B_m$ is a clause of the program, the component atom A_i may be proved by proving each of B_1, \dots, B_m and the goal A_1, \dots, A_n may be reduced to one in which A_i is replaced with B_1, \dots, B_m .

The mechanism for answering queries is somewhat more complicated when variables are allowed to occur in a query or in the clauses of the program. In this case, the query $?A_1, \dots, A_n$ asks which individuals satisfy $A_1 \& \dots \& A_n$, and the corresponding goal is to prove an instance of $A_1 \& \dots \& A_n$, i.e. to prove each of A'_1, \dots, A'_n for some instance A'_1, \dots, A'_n of A_1, \dots, A_n . Accordingly, in reducing one goal to another, we need no longer require that a component atom of the query exactly match the head of a clause; we can merely require that the component atom and head have a common instance. The replacement can then be made on that common instance. Suppose, for example, that the clause is $Pfby \leftarrow Ry$ and that the goal is $Pfxgz, Qx$. Then the substitution $\theta = \{x/b, y/gz\}$ (taking x to b and y to gz) "unifies" the atom $Pfxgz$ and the head $Pfby$, it yields a common instance $Pfbgz$; and so $(Pfxgz)\theta$ in the instance $(Pfxgz)\theta, (Qx)\theta$ of our original goal may be replaced with $(Ry)\theta = Rgz$ to produce the new goal Rgz, Qb .

In making this transition from one goal to the next, it is helpful if two restrictions are observed. First, the variables of the clause should be rewritten so as to be distinct from those in the query. Second, the substitution should be selected so as to be a most general unifier (*mgu*) of the head of the clause and the component atom. Any other unifier of the head and the atom should be a refinement of the given unifier. These two restrictions guarantee that the reduction is of the most general form; and no other unifier of the head and atom need therefore be considered.

How is this mechanism for reducing one goal to another used to answer

a query $?Q$. The mechanism may be repeatedly applied to obtain a series of goals $G_0, \dots, G_k, k \geq 0$. Suppose now that the last of these goals G_k is empty. Then since this goal evidently succeeds (each member of an empty set of literals is provable), our original goal must succeed and so the given query $?Q$ must have a positive answer. In case $?Q$ is ground, that positive answer can only be "Yes". But in case $?Q$ is unground, the answer is a specification of values for the variables of Q ; and it may be obtained from the substitutions $\theta_1, \dots, \theta_k$ by which the reductions are made by restricting their composition to the variables of Q .

Thus the positive answers to a query are obtained from the successful reduction paths, those that terminate in the empty goal. But how are the successful reduction paths to be found? One possibility is to lay out a space of reduction paths in the form of a tree, the so-called *evaluation tree*. Each node is labelled with a goal. An atom is, if possible, selected from the goal. The descendants of the node are then labelled with the reductions on that atom. Thus paths of the tree correspond to reduction paths; and the successful reduction paths may be obtained by searching the tree in any one of a standard number of ways. It is usual in Prolog to select the left-most atom and to make a depth-first search of the tree; but we shall not be concerned with either the questions or the difficulties which arise from adopting different selection procedures or search strategies.

We have shown how positive answers to a query can be obtained. However, Prolog also provides a method for obtaining a negative answer "No"; this is the rule of negation as failure. It is this rule that has been found so problematic and provides the focal point for the present paper. Let us say that a node *fails* if its goal is non-empty and cannot be further reduced, that a path of a tree *fails* if it terminates in a node that fails, and that the tree itself *fails* if each of its paths fail. A query or goal is said to *finitely fail* (*ffail*, for short) if an evaluation tree for it fails. Thus a goal *ffails* if a systematic attempt to achieve it, or prove what it states, breaks down in a finite number of steps. The rule NF of negation as failure then allows us to return the answer "No" to a query when it *ffails*. In the program $\{p \leftarrow q, q \leftarrow r, q \leftarrow s\}$, for example, the answer to the query $?p$ will be "No", since each of the two attempts to prove p (via $q \leftarrow r$ and $q \leftarrow s$) will terminate in failure.

The rule of negation as failure has so far been applied externally to the reduction process as a means for producing negative answers, but it can also be applied internally to the process as a means of reducing one goal to another. Suppose that $\neg A$ is a component literal of a goal. Then if the

simple goal A fails, we may reduce the original goal to the result of removing the component literal from it. The case in which A succeeds also provides us with a new means for identifying failure. For we may then take any node labelled with the original goal to fail. (If the atom A is unground, then it should actually be required that A succeed with the identity substitution or a trivial variant thereof.)

With such an extension, negation can sensibly be used in queries and in the bodies of program clauses. Consider, for example, the program $\{q \leftarrow \neg p\}$. Then the goal q will reduce to $\neg p$. But p fails; and so $\neg p$ reduces to the empty goal and the original goal succeeds.

Of course, if a negative literal is selected from a goal, it must somehow be ascertained whether it fails or succeeds. This requires the search of a further tree and yet further trees still, should these be invoked by the selection of other negative literals. Thus the search through a single tree must be replaced by a search over a family of trees. Again, we shall not explore the questions or difficulties which arise from the different ways in which this might be done.

The rule NF has been stated as a rule of reduction. It can equally well be stated, in more orthodox form, as a rule of inference. In the case in which variables are excluded from the language, the rule allows us to infer the single literal $\neg B$ from $\neg A_1, \dots, \neg A_n$, under the condition that each clause whose head is B contains one of the literals A_1, \dots, A_n in its body. In the case in which variables are permitted, the condition is more complicated. We require that each instance of a clause whose head is an instance of B should contain an instance of one of the literals A_1, \dots, A_n in its body. The rule can be further extended to the case in which a conjunction of literals is to be denied; and combining such an extension with classical rules allows one to present Prolog as an inferential system of a standard kind.

2. What is the problem?

Before discussing the specific solutions to the problem of justifying the rule of negation as failure, it will be helpful to consider the problem itself in a more general light. It will be important, in particular, to distinguish among the different situations in which the demand for a justification might be made and to evaluate the different requirements that might be imposed on the way these demands are to be met.

The problem of justifying the rule of negation as failure only sensibly

arises in relation to an interpretation for the language of the program. For what we want to know is whether the answers provided by Prolog are correct. But without an interpretation, there is no concept of truth; and without a concept of truth, there is no question of whether an answer is correct.

Let us use the term *situation* for an ordered pair (M, P) consisting of a program P and a model M for the language of P . Thus situations in the technical sense correspond to situations in which a programmer actually finds himself. For in writing a program, he will usually have a particular interpretation in mind, one to which the program is meant to conform.

There may also be situations in which the programmer has several interpretations in mind. He may be interested, for example, in the truths of some particular algebraic theory. In these cases, it is still by reference to the intended interpretations that the concept of a correct answer is to be understood; for what renders an answer correct is that it should be true under all of the intended interpretations. The corresponding technical notion of situation must now be of an ordered pair consisting of a program and a *class* of models for the language of the program.

In what follows, we shall confine our attention to those situations in which the program is true in the intended model (or models). This *appears* to be eminently reasonable; for how can the answers provided by Prolog be correct when the program itself is not. However, this restriction is not quite as compelling as the reason given for it would seem to suggest.

It is true that, for most systems of reasoning, the conclusion that can be drawn from a set of premises will have the premises themselves as logical consequences. Thus any infirmity in the premises must show up in the conclusions; if the premises are not all true then neither are the conclusions. However, this is not a feature of the Prolog system of reasoning. Of course, we cannot directly ask of a program clause whether it is true, at least in the usual versions of Prolog; for it is not of the right form. The program clauses will not in general even be logical consequences of the answers to the queries that can be posed. Consider, for example, a situation in which the program is $\{q \leftarrow p, q \leftarrow q, p \leftarrow p\}$ and the intended interpretation makes p true and q false. Then the program is false and yet the program delivers no answers and, *a fortiori*, no false answers.

There might even be some instrumental value in programming with falsehoods; for perhaps this is more efficient or more economical than programming exclusively with truths. One would need to be confident that in such cases there was no spill-over of falsehood from the program

to its consequences. For this reason, and also because of its intrinsic interest, it would be desirable to have a better technical understanding of the situations in which the program can be false and the conclusions true. However, in the present paper, I shall not consider such situations: partly because the programmer usually strives to construct a true program (even if he does not always succeed); and partly because the special problems raised by false programs are best considered separately.

Even with this restriction, the nature of the problem of justification very much depends upon the range of situations to which the rule is to be applied; for a justification which is satisfactory for one range of application may not be satisfactory for another. In this regard, there are two distinctions among situations that are of great importance. (Various technical conditions on programs and queries have been proposed in the literature that are relevant to the applicability of different justifications and the agreement among them. The present distinctions are of a more elementary and fundamental kind, and concern, not the programs themselves, but their relationship to the intended model or models.)

There is, first of all, the distinction between those situations in which the failed statements are all false and those in which they are not. The significance of the distinction is this. In considering the former kind of situation, it is possible to interpret negation classically, i.e. in accordance with the precept that a negative statement $\neg A$ is true iff the negated statement is false; for each failed statement is false and so its negation, which is yielded by the rule NF, is true. It is therefore possible, at least in principle, to find a justification of the rule which respects the classical reading of negation; the justification can be semantically conservative. However, in the latter kind of situation, it is not possible to interpret negation classically; for some failed statement is, by supposition, true and so its negation is false. If therefore all of the Prolog consequences are to be true, some other way to interpret negation must be found; the justification must be semantically innovative.

There are two main factors which are relevant to our confidence in the falsehood of the failures and which are therefore relevant to the question of which kind of justification is available to us. The first is the extent to which there are failures; for the more there are, the harder it will be to know that they are all false. The second is the extent to which there are known falsehoods; for the more there are, the easier it will be to know that they include the failures. (I talk of knowledge; but similar points apply, both here and elsewhere, with comparable epistemic attitudes in its place.)

The first factor depends upon how the program is written. To take an extreme example, if clauses $A \leftarrow A$ are added for each simple atomic statement A then there will evidently be no failures and, on that basis alone, we can ascertain that they are all false. The second factor depends, of course, upon our knowledge of the subject-matter. As a general rule, this will be quite adequate in the case of mathematical domains, for we will know of each ground atom whether it is true or false; but it may be more or less inadequate in the case of empirical domains. (Two subtle, yet important, qualifications should be made. If the failures are allowed to be unground, then our knowledge of general falsehoods, which is usually much more problematic than our knowledge of particular falsehoods, must also be considered. But even if the failures are required to be ground, one may know that each failure is false without knowing that every failure is false. To know the latter claim may require a general mathematical understanding of the program which is also hard to come by.)

The other distinction that needs to be drawn is between those possible situations in which the rule NF *might* correctly be applied and those actual situations in which the rule NF *is* correctly applied. In talking of actual situations, I have in mind those that are of the kind in which the programmer finds himself. I take it that among those situations there are some, of a characteristic kind, to which the rule NF is correctly taken to apply. In talking further of correct application, I mean to restrict the actual situations to those which are of this characteristic kind.

It would be idle to pretend that we possess a sharply defined concept of characteristic situation. The point is that the situations in which we write programs are significantly different from situations in general. We are not merely writing down truths, but are also trying to meet certain computational demands; and as anyone who has used Prolog will know, these demands very much shape what it is that one writes down.

We may attempt to justify the rule NF either in its application to the possible or to the actual situations. The bearing of this distinction on the question of the justification might appear to be evident. For in so far as the intended application of the rule is less, then less will be required of a justification and therefore the greater will be the chance that a justification is found.

This difference would be significant if there existed no broader justification. But what if there did? Even so, the narrower justification might be more satisfactory in certain respects. But what if the broader justification were as satisfactory as the narrower justification? Why not simply go for

the most satisfactory justification with the broadest range of application?

I think this would be the right strategy if our only concern were to justify the rule of negation as failure. However, it seems to me that we have special reasons for restricting our attention to the narrower class of situations. For our interest is not only in justifying the rule in the situations in which it is characteristically applied, but also in determining what other rules might legitimately be applied in those situations. If we merely say what it is about those situations that entitles us to use the rule NF, then we may not have said enough to justify the application of those other rules.

In general, it seems to me that there are two quite different justificatory tasks that we may set for ourselves. On the one hand, our focus may be on the rule NF itself. Keeping this fixed, we may aim for breadth and attempt to find the widest range of situations in which the rule might justifiably be applied. On the other hand, the focus may be on the situations in which the rule is characteristically applied. Keeping them fixed, we may aim for depth and attempt to find the widest range of rules which can justifiably be applied within them. In the first case, we are interested in what it is about the rule that might justify its application within the characteristic situations and, one hopes, in other situations as well. In the second case, we are interested in what it is about the characteristic situations that might justify the application of the rule NF and, one hopes, other rules as well.

I do not wish to dispute the interest of the rule-oriented problem, but it is easy to exaggerate its significance. For it is not as if we have envisaged any extension of the rule NF beyond the range of situations in which it is normally applied. Any new cases considered would therefore appear to be more of theoretical interest and to serve more the purpose of illuminating the nature of the rule rather than of grounding our inferential practice.

By contrast, the situation-oriented problem is of direct relevance to an envisaged extension of our inferential practice. For we wish to know whether any further rules of inference are justified in those situations in which NF is characteristically applied. This question is, in its turn, relevant to the question of whether Prolog should be supplemented with further non-classical rules besides NF. Of course, any proposed rule might be too difficult to implement or to implement in full, but we should at least decide whether we want a rule before attempting to determine whether we can have it. Later I shall suggest that there are indeed some further non-classical rules that might legitimately be added to Prolog.

I have so far talked of what a justification should justify, but I have not attempted to explain how a justification should justify. What is it that makes one justification satisfactory and another not? One obvious formal requirement on a proposed justification is that it should be one for which the rule NF is sound: whenever the given situation has whatever feature is enjoined by the justification, then any statement inferred by means of the rule should be true.

However, it is not clear to me that there are any other formal requirements that might reasonably be imposed upon a justification. Two others have sometimes been suggested. The first is that the justification should be one for which the rule, in the context of Prolog, is complete. However, there are several different things that might be meant by completeness here; and it will be worthwhile to distinguish them, partly because of their intrinsic interest and partly because of their bearing on the present question.

There is, first of all, an interrogative concept of completeness. This is purely internal to the system and makes no reference to an external standard of correctness. Roughly speaking, a query-answering system is complete in this sense if it does not matter which queries you pose. To be more exact, it must be true for any queries $?Q$ and $?R$ that if an answer-substitution σ for $?R$ provides an answer-statement for $?Q$ (i.e. $R\sigma$ is an instance of Q), then some answer-substitution θ for $?Q$ must provide an equally good answer-statement for $?Q$ (i.e. $R\sigma$ is an instance of $Q\theta$). It is clear that if a system is incomplete in this sense then this must be because of a defect in the ability of the system to answer queries and not a defect in any justification or external standard of correctness that might be provided.

There is, secondly, the declarative concept of completeness. This is the standard concept, familiar from the study of logic. A system is complete in this sense if every correct statement of the language is derivable. The derivable statements of Prolog are those that are determined by the answers that it provides. It is usual to combine the concepts of interrogative and declarative completeness, but it is clearly preferable to isolate a purely internal aspect of completeness, as I have done here.

The declarative concept of completeness may itself be broken down into two aspects. There is, in the first place, the ability of the system to draw classical consequences (or to draw logical consequences should some other form of logic be adopted). The system is complete in this sense if any statement (of an appropriate form) which is a logical consequence of the program and the derivable statements is itself deriv-

able. Note that this is stronger than the condition that any statement (of the appropriate form) that is a logical consequence of the program should be derivable, since the system may not be logically sound. It is again clear that if the system is incomplete in this sense then this must be because of a defect in the ability of the system to draw logical consequences and not because of any defect in the justification.

There is, in the second place, the ability of the system to draw non-classical (or non-logical) consequences. This gives what one might call a *supra-classical* (or *-logical*) concept of completeness. A system will be complete in this sense if any correct statement (of the appropriate form) is a classical (or logical) consequence of the program and the derivable statements. It seems that if a defect in a justification is attributable to incompleteness, then it must be incompleteness in this, and not in any of the other, senses.

Whether supra-logical incompleteness should be regarded as a defect in a justification will depend very much on the nature of the justificatory task. If we are aiming for breadth of application, then it would appear to be a defect. For if some correct statement (of the appropriate form) which is deemed correct under the justification is not a logical consequence of the program and the statements derivable from the program, then there is a model in which the program and derivable statements are true and the putatively correct statement is false. Since the program and the derivable statements are true in the model, there is a *prima facie* case that the situation comprised of the program and the model should be covered by a comprehensive justification; but since the putatively correct statement is false in the model, it will not be covered by the given justification.

On the other hand, if we aim for depth, and focus our attention on the situations in which the rule NF is characteristically used, then it is not even reasonable to expect that the system should be supra-complete. For the rules which are added to Prolog are very much constrained by considerations of implementation; and so even if we supposed that we had done as good a job as possible in implementing the non-classical rules and even if we supposed that in this case, as in others, the correct non-classical inferences could *in principle* be formalized, there would still be no reason to suppose that they had actually been implemented.

Another formal requirement that has been imposed on a justification is that it should lead to a computationally tractable notion of correct consequence (I suppose that it is thought that the more tractable the notion the better the justification). However, it is not clear to me why this

requirement has been thought to be so compelling. There are philosophical reasons for insisting upon a constructive justification, one that employs constructive concepts and constructive principles of reasoning, but I take it that these reasons are not at issue here. Therefore, why should it not be as legitimate to appeal to a computationally intractable standard of correctness in this case as in the case, let us say, of arithmetical truth?

SHEPHERDSON (1988) has suggested that if the concept of correctness were no more tractable than the concept of first-order validity then we might as well aim to implement the concept of first-order validity. Of course, this makes computational tractability not so much a requirement on a justification of the rule NF as a reason for adopting the rule in the first place. The thought seems to be that we should try to implement the more implementable standard of correctness: it is "easier"; and we can get "closer" to our goal (not that we can succeed in either case). But it may be better to aim high than low, for we may have a greater interest in what is thereby achieved. Thus the fact that we have chosen to implement a non-classical rather than a classical concept of correctness would appear to suggest that we have a greater interest in the statements that are obtained under the one approach than under the other; and this is an interest that remains regardless of how relatively untractable the standards of correctness might be.

In general, there seems to be something misguided in aiming for a computationally congenial standard of correctness; it is almost as if one were to aim for a decidable notion of first-order validity. There are other considerations altogether which determine whether a standard of correctness is satisfactory. However, once a standard is found, its computational properties can be investigated and an attempt can be made to isolate tractable sub-classes of interest should the whole class of correct statements prove to be intractable.

Although it seems to me there is only one formal requirement, that of soundness, which can reasonably be imposed upon a justification, there is also a very important informal requirement, which I call "availability". This is of an epistemological character and is roughly to the effect that the justification should put us in a position to use the rule. The need for this requirement can be made evident by asking: why is it not satisfactory to justify an application of the rule NF on the grounds that all of the failures are false? I take it that the reason is that the justification does not provide us with the grounds upon which it is proper to believe that the failures are false. In other words, a justification should make it clear *how* we are in a position to know that the failures are false in those

situations in which we are in a position to know that the failures are false. The justification should provide us with an epistemic access to the proper application of the rule.

It is important to distinguish between the present requirement of availability and the previous requirement of computational tractability. One may be in a good position to know that the consequences of a program or theory are correct even though the standard of correctness is far from tractable. After all, one knows that the classical consequences of Peano Arithmetic are correct under the highly intractable standard of arithmetical truth. Whether a proposed justification meets the present requirement depends, not on a proof that it possesses computationally desirable properties, but on an informal consideration of the kinds of argument that are available to us in showing that the demands of the justification are met in a given case.

This requirement has largely been ignored; perhaps because the interest has been in more formal criteria. However, it should be clear that such a requirement is at the heart of the concept of justification; and, indeed, we shall later find it of great value in evaluating various proposals that have been made.

3. Negation re-interpreted

If we wish to adopt a rule of inference that is not classically valid (or reject a rule that is classically valid), then one way to justify the adoption (or the rejection) is to reinterpret the logical constants. This is the way, for example, in which one might attempt to justify the adoption of the Aristotelian rules of the syllogism (in which universal statements have existential import), or to justify the constructivist's rejection of the rule of double negation (by which A can be inferred from $--A$). In the present case, the rule of negation as failure is invalid under a classical reading of negation (and, one should add, of the conditional and conjunction). It is therefore only natural to consider whether the rule might be valid under a non-classical reading of the connective.

Such a reading comes immediately to mind. Interpret the negative statement $\neg A$ as simply the statement that A finitely fails. Each statement inferred with the help of the rule of negation as failure will then be true. It will not be quite right to say, however, that the inferred statement actually follows, under this reading, from the clauses of the program

themselves; it follows rather from the fact that they are the clauses of the program.

The interpretation of negation as finite failure is the most informative that can be given (at least on ground atoms), but it is not by any means the most interesting. Indeed, it is not clear that any intrinsic interest attaches to the concept of finite failure at all. However, the fact that a statement fails may imply, either on its own or in the presence of further assumptions, that it has certain features of interest; and these may also be used to interpret negation.

Thus from the failure of a statement it follows that it is underivable, both in classical logic and in Prolog; and so it is possible to interpret negation as underivability in either of these two senses. It is natural to make the supposition that a statement and its negation cannot both be true (though it is not actually necessary for a justification of NF to succeed). Under this supposition, unprovability in Prolog would then constitute the broadest, i.e. the least informative, re-interpretation of negation.

Provability may be related to knowledge. Granted that the program embodies our knowledge in the sense that every known truth is derivable (either in Prolog or in classical logic), it follows that every unprovable statement is unknown; and consequently negation can be interpreted as "not known to be true". Granted also that each derivable statement is known, the interpretation will be acceptable in the sense that no statement and its negation will both be true.

There is no doubt that there are situations in which one is forced to adopt something like this justification of the rule. Of course, some of the failures may be true; and so that is the end of the matter. However, even when all of the failures are false, one may be in no position to know that they are all false; and so a re-reading of negation will have to be made if the justification is to be available.

However, there are also situations in which one does feel justified in adopting a classical reading of negation. The clearest cases are those in which the program takes the form of an inductive definition of a mathematical predicate. Perhaps I have set up the following program for the predicate of being even: $\{Eo \leftarrow, Essx \leftarrow Ex\}$. I am then in no doubt that the finitely failed statements of the form $Es \dots so$ are false. But the present approach could only explain why I am justified in believing them to be unprovable or unknown.

There is a related difficulty. This concerns not the external use of negation in framing answers, but the internal use in framing clauses. If I

interpret negation as finite failure, let us say, then I can no longer have the assurance, which I would expect to have, that the clauses of my program are correct. I may add to my program for the predicate of being even a clause for the predicate of being odd: $Ox \leftarrow \neg Ex$. Under a classical reading of negation, there is no difficulty in seeing that this clause is correct, but with negation read as finite failure, seeing that such clauses are correct will usually require much more insight into what fails than I possess. The difficulty becomes even more acute under the broader (and usually more interesting) readings of negation, since then the negative clauses become stronger in content.

One might try interpreting the negation of clauses classically and the negation of answers proof-theoretically or epistemically, but one would then no longer be justified in using negative answers to detach the negative antecedents of clauses. It therefore seems that the re-interpretations of negation that we have considered will have very little useful application to programs containing negative clauses.

4. Closed worlds

I wish now to confine my attention to the case in which the rule NF is to be justified under a classical reading of negation (and of the other logical constants). There are two main ways in which such a justification can be presented. The first is syntactic and consists in providing a transformation P' of any program P (or of any program P from a given range). For the most part, P' will be the result of adding certain "tacit" assumptions to P , although other forms of transformation can in principle be countenanced. It is supposed that by the program P the programmer really means P' . The justification of the rule NF then rests on the fact that any statement which can be inferred from the program P by means of Prolog can be inferred from the transformed program P' by means of classical logic. Of course, some other logic can be used in place of classical logic to effect the inferences from P' , as long as its own justification is not in question.

The other way of presenting a justification is more semantical in style. The truth of a program P in a model M (under the classical truth-conditions) is not sufficient to guarantee the truth of those statements which can be inferred from P by means of the rule NF. Therefore a closer, or more intimate, relationship than mere truth is required to obtain between a program and a model if the truth of the inferred

statements is to be guaranteed. The semantical-style of justification consists in specifying such a relationship.

For the most part, this more intimate relationship (call it \models') will be explained as a refinement of mere truth (which we denote by \models); \models' will be \models plus something more. But it is not ruled out that \models' might, in principle, be explained in some other way. It is now supposed that when the programmer asserts a program P , he not only wishes to assert that it is true in the intended model (or in the intended models) but also that it is true or correct in this more special way. The justification of the rule NF then rests on the fact that any statement which is inferred from the program P by means of the rule will be true in the (or an) intended model as long as the program bears this more intimate relationship to the model.

I talk of two different forms of presentation rather than of two different types of justification, since it will always be possible to present a given justification in either of the two ways. For given a syntactical transformation P' , we can define a semantical relation \models' by the condition: $M \models' P$ iff $M \models P'$; and given a semantical relation \models' , we can define a syntactical transformation P' by the equation: $P' = \{A : M \models A \text{ whenever } M \models' P\}$. Some information may be lost in the passage from one notion to the other, but whichever notion is taken as primitive, it will be true that:

$$P' \vdash A \text{ iff } M \models A \text{ whenever } M \models' P.$$

Of course, if one style of justification is obtained from the other in this mechanical manner, then the definition of P' may be semantical in character and the definition of \models' may be syntactical in character. However, in most of the cases of interest, it will be possible to provide independent specifications of P' and \models' . Thus it will be genuinely illuminating that the justification has both a syntactic and a semantic formulation.

Our first approach to justifying the rule NF under a classical reading of negation is provided by the closed world assumption of REITER (1978). This assumption is usually presented in the form of a rule:

If the positive statement A is not derivable, then the negative statement $\neg A$ may be inferred.

There are, however, two other items that might legitimately be regarded as the closed-world assumption. The first is a single *statement* to the effect

that $\neg A$ holds for every positive statement A not derivable from the given program P . This is a logically complex statement. It may be formulated in different ways, but it will usually call for resources beyond those of a first-order language. The assumption might also be taken to be the set of negative statements $\neg A$ for which A is not derivable. Each member of the set will, of course, be of a simple form; but the set itself may be infinite. It will usually be apparent from the context which of these three items — the rule, the statement or the set — is meant by the closed-world assumption; and therefore I shall not be scrupulous in carefully distinguishing between them.

The closed-world assumption probably constitutes the most immediately attractive approach to the problem of justifying the rule of negation as failure. For the simplest case in which the rule is applied is to a list of relational statements, as in a time-table or catalogue. In this case, it is merely the presence or absence of a fact from the list which indicates whether it is to be asserted or denied. However, when we add logically more complicated statements to the list, such as unground or conditional clauses, we can no longer appeal to simple presence or absence, since a statement which is absent from the list may be derivable from statements which are present in the list. The most natural extension of the idea is therefore to substitute the notions of implicit presence and absence, i.e. of derivability and underderivability, for the notions of explicit presence and absence. But this amounts to no more than the closed-world assumption: statements that are not derivable are to be denied.

The assumption, for all its appeal, is somewhat indeterminate, since it needs to be said what is meant by “derivable”. It is most natural, in the present context, to take “derivable” to mean “derivable within classical logic”. It is also possible to take it to mean “derivable within Prolog”. Later we shall provide another, rather different, understanding of derivability.

If derivability is taken to be within classical logic, then the closed-world assumption corresponds to the requirement that an intended (classical) model should make only the derivable positive statements true. Seen as a condition on the program, this amounts to the familiar claim of completeness (though under a restriction to positive statements): no truth is underived. Seen as condition on the model, it amounts to a claim of minimality: no underived statement is verified.

In its application to positive programs (i.e. to programs without negation), the requirement has an illuminating independent formulation in terms of inductive definition. A positive program may be regarded as a

means for generating positive statements. For the categorical clauses generate positive statements directly; and the conditional clauses generate positive statements indirectly, from others. A model is taken to be *inductively defined* by a program if the set of its positive truths coincides with the positive statements generated by the program. If we say that a model for a program is a *fixed point* when every truth within the model is generated from a truth, then the inductively defined models will be the least of the fixed points, i.e. the fixed points whose class of truths is included in all others. (It should be noted that the usual account of inductive definition is somewhat more general, since it determines the satisfaction of the formulas $Px_1 \dots x_n$, upon the basis of a prior specification of the interpretation of the function symbols and constants, and not merely the truth of the sentences $Pt_1 \dots t_n$. On the present account, certain difficulties arise if the model is allowed to contain undesigned individuals. These will be discussed later.)

The inductively defined models, as so characterized, are exactly the models that conform to the closed-world assumption. Indeed, the assumption is merely the counterpart (though for the generation of truths, not extensions) of the familiar extremal clause. Let us say that a model is *implicitly defined* by a program (or definition) if it renders the program (or definition) true. The closed-world assumption, or the extremal clause, can then be regarded as a device for converting an inductive into an implicit definition. A model will be inductively defined by a program just in case it is implicitly defined by the result of adding the closed-world assumption to the program.

The great strength of the closed-world assumption lies in its application to positive programs, those that make no use of negation in the body of their clauses. For positive programs are characteristically used to generate the positive truths. The intended model is inductively defined by the program; and the application of the rule NF is thereby justified.

Nor should it come as a surprise that positive programs characteristically have this inductive role. For the situations in which we frame programs are of a very special sort. We are not merely concerned to write down truths; we also want to meet certain computational demands. These are quite reasonably taken to require that every (positive) truth should be derivable within Prolog from the clauses of the program; and so they will guarantee that the clauses will indeed constitute an inductive definition.

The weakness of the closed-world assumption lies, by contrast, in its application to negative programs, those that do make use of negation in the body of clauses. For the assumption (in its syntactic formulation) will

often render such programs inconsistent, and it is therefore impossible that the assumption (in its semantic formulation) should hold. Consider, for example, the program $\{q \leftarrow \neg p\}$. Then neither q nor p is a classical consequence of $q \leftarrow \neg p$. Thus, $\neg q$ and $\neg p$ should be added to the program, which is thereby rendered inconsistent. (If consequence within Prolog is used in place of classical consequence in the formulation of the closed-world assumption, then the clause $p \leftarrow p$ may be added to the program for the example to work).

This example would be of no great significance if it were not typical of one of the ways that negation is used in the formulation of Prolog programs. For our aim is, or at least may well be, to justify only the actual use of the rule of negation as failure, not every possible use. Unfortunately, it is very common for negation to be used in just the way required by the example. I might wish to add, for instance, a clause $Ox \leftarrow \neg Ex$ for the predicate of being odd to a previous program for the predicate of being even. This clause would then lead to exactly the same difficulties as the earlier clause $q \leftarrow \neg p$.

A related difficulty concerns the issue of availability or epistemic access. We want a justification to explain how we can know that the application of NF is sound and not merely to take that knowledge for granted. In the case of positive programs, the approach does well; for we can often appeal to our inductive understanding of the structure at hand to show that every truth is generable. Thus there is no difficulty in showing that the program $\{Eo \leftarrow, Essx \leftarrow Ex\}$ for the predicate of being even generates exactly the positive truths; for we can give a simple inductive proof, on the basis of the construction of the term t , that every truth of the form Et is generable.

In the case of negative programs, the approach fares much worse. Even should the truths be exactly the consequences of the program, we will rarely be in a position to show that they are; for there is no straightforward inductive argument and, in general, no comparable argument that will enable us to see how use might be made of the negative clauses in the program. Thus the proposal will leave us very much in the dark as to the application of the rule NF in such cases.

5. Completed domains

The third approach to justifying the rule of negation as failure is provided by the account of CLARK (1978) in terms of *completed domains*.

Again, this account has both a syntactic and a semantic side. Under the syntactic formulation, it is tantamount to treating the sufficient conditions given in a program for the application of a predicate as jointly necessary. Thus, the simple program $\{q \leftarrow p\}$ (with 0-place “predicates” q and p) is taken to be equivalent to $\{p \leftrightarrow \perp, q \leftrightarrow p\}$. On the semantic formulation, it amounts to the assumption that the intended model is a fixed point: a positive statement is true if it can be generated by means of the program from other truths (either positive or negative).

Thus we see that both Clark’s and Reiter’s accounts amount to adopting an implicit assumption of completeness, an assumption that is very natural once some form of non-monotonic reasoning is in question. However, they each take a very different view of what that completeness consists of. For Reiter it is completeness with respect to the derivability of truth, all truths must in effect be given; for Clark it is completeness with respect to the specification of sufficient conditions, all sufficient conditions must in effect be given.

Clark’s account appears to perform better on negative programs than Reiter’s. Consider the previous example of the troublesome program $\{q \leftarrow -p\}$. Reiter’s account renders the program inconsistent. Clark’s account, on the other hand, renders it equivalent to $\{q \leftrightarrow -p, p \leftrightarrow \perp\}$. This is, in its turn, equivalent to $\{q, -p\}$, which is just what the application of the rule NF requires.

Clark’s account does indeed render some programs inconsistent. The simplest example is $\{p \leftarrow -p\}$, which converts to $\{p \leftrightarrow -p\}$. However, in contrast to the case of $\{q \leftarrow -p\}$, the present case is not typical of the kind of program that anyone would normally want to use. We never, or never should, write a program in such a way that the truth of a statement can only be generated from its falsehood. For either the statement is false, in which case we are also required to take it to be true; or else it is true, in which case we are given no means for its generation.

This is, of course, just one example; other cases are much less blatant. However, as far as I can see, the point generalizes. Clark’s account never delivers more than we would want to accept, even for negative programs, in the standard situations in which programs are used.

However, Clark’s proposal does sometimes deliver *less* than we would want to accept; and this is true even for the case of positive programs. For his proposal only enjoins us to accept the falsehoods of the greatest fixed point; and yet it is characteristically the falsehoods of the least fixed point that we are willing to accept. In writing a program, we are trying to meet a certain computational demand, viz. that all positive truths be

derivable. However, as I have already pointed out, this demand cannot be met unless the intended model is a least fixed point for the program. Of course, it may be that, in a given case, the demand cannot be met or, in its full extent, is not required. Then we would simply have a computationally defective program; and there would be no reason in general for supposing that it was defective in the peculiar way required for the intended model to remain a fixed point. Clark's proposal misrepresents the programmer as trying to write a program for which the intended model will be some fixed point or another; whereas it is only the least fixed points for which he aims, if he aims for any fixed point at all. If he ends up with some other fixed point, it will be by accident rather than by design.

There is another approach to the completed domain account, which uses a 3-valued semantics and logic in place of the classical semantics and logic adopted by Clark (see FITTING 1986, KUNEN 1987). If we are aiming for breadth of application, then it is certainly of interest that the justification of the rule is not tied to a 2-valued semantics or logic (and this is an interest which would remain even if the 2-valued semantics provided a completeness result for Prolog). However, if our concern is with what I have called depth, then it is not clear that the 3-valued is any better than Clark's original 2-valued approach. For what we then require of a justification is that it should show how the statements inferred by means of the rule are correct under the *intended* interpretation (or interpretations). However, as a rule, the programmer will have a 2-valued model (or models) in mind; and even when the model admits gaps of a normal sort, perhaps through vagueness or empty reference, there is no reason to suppose that the gaps will so arrange themselves as to render the biconditionals true.

It is often said that the 3-valued semantics arise very naturally within the context of programming languages; for the evaluation of a statement may either lead to the answer "True" or to the answer "False" or to no answer at all. This is so. But such an interpretation merely provides us with a meta-logical interpretation of the simple predicates of the language, and so the clauses of the program end up as saying something about the computational properties of the program itself. However, the programmer is not interested in the computational properties of the program as such, but only in so far as they bear on real questions of truth and falsehood. He may want to know what is computably true or false, but only because he believes that what is computably true or false is actually true or false. Thus it is only to the extent that the 3-valued

interpretation can be related in this way to a real 2-valued interpretation that the question of justification can sensibly arise.

The 3-valued approach was no doubt inspired by the allure of completeness (not that it has been altogether successful in this regard), but the superior scope of the 3-valued approach is attributable solely to the classical deficiencies of Prolog. It is only, for example, because Prolog does not permit the classical inference of p from the program $\{p \leftarrow \neg p\}$ that the 3-valued approach is able to account for the non-failure of p , in contrast to the 2-valued approach. Repair the classical deficiencies of Prolog and the supposed superiority of the 3-valued approach disappears.

If we insist on retaining the intended classical interpretation of Prolog, then the real question of interest becomes, not the completeness of Prolog with respect to an expanded range of non-classical models, but the supra-classical completeness of Prolog with respect to a limited range of classical models. We wish to know, not what strange models might justify the defects in Prolog, but what restrictions on the standard models might justify its excesses.

If these criticisms of the standard proposals for justifying the rule of negation as failure are accepted, then they leave us in a predicament. For it seems that we must either accept the closed-world assumption, which is too strong for negative programs, or else accept the account in terms of completed domains, which is too weak for positive programs. The problem is to find justification which strikes the proper balance between these two accounts.

6. Self-referential closed worlds

I shall now present my own proposal, which may be viewed as a way out of the predicament. Any program, positive or negative, can be viewed as a mechanism for generating truths: for we are told outright that certain truths obtain; and we are also told that if certain truths obtain then so do others. Now from this point of view, program clauses containing negation in their body are useless. For we never generate falsehoods, and so we are never in a position to detach all of the antecedents that figure in the body of such clauses (or of their instances).

In advance of the generation procedure, suppose that we make a hypothesis as to which statements are false. Then this hypothesis can be used, as part of the procedure, to detach negative statements from the bodies of clauses. Therefore, if we are given the clause $r \leftarrow p, \neg q$, for

example, and if the truth p has already been generated, then the hypothesis that q is false can be used to generate the truth of r .

If an hypothesis as to which statements are false is used in this way to generate truths, then there are three possible outcomes: some statement is neither a posited falsehood nor a generated truth (a “gap”); some statement is both a posited falsehood and a generated truth (a “glut”); the posited falsehoods are the exact complement of the generated truths (no gap and no glut). Suppose, for example, that the program is $\{q \leftarrow -p\}$. Then the hypothesis that posits no falsehood leads to a gap for p and q , since no truths can be generated; the hypothesis that posits the falsehood of both p and q leads to a glut since, given that p is false, q is a generated truth; while the hypothesis that posits the falsehood of p alone leads to neither gaps nor gluts, since q is the sole generated truth.

Let us call a hypothesis *happy* if it leads to neither gaps nor gluts. Instead of thinking in terms of posited falsehoods, we can think in terms of posited truths. A happy hypothesis is then one under which the posited truths coincide with the generated truths. Thus a happy hypothesis is, in a certain sense, self-verifying. It is *verifying*, since what one takes to be the truth turns out to be the truth; and it is *self-verifying*, since it is partly because one takes the truth to be what it is that it is what it is.

In the program $\{q \leftarrow -p\}$ above, there is just one happy hypothesis. For the only hypothesis not considered is the one that posits the falsehood of q alone; and this leads to a gap for p . Thus, of the four possible hypotheses, only “ p false” is happy. However, in general there may exist no happy hypotheses or several. The program $\{p \leftarrow -p\}$ has no happy hypothesis. For p false leads to p true and hence a glut; while p not false leads to nothing and hence a gap. On the other hand, the program $\{q \leftarrow -p, p \leftarrow -q\}$ has two happy hypotheses. For p false leads to q true and q false leads to p true; while p and q both false or both not false fail for the same reasons as before. (We may note that the present program is logically equivalent to the previous program $\{q \leftarrow -p\}$ with one happy hypothesis; and so the happiness of a hypothesis need not be preserved under logical equivalence.)

Let us call a model for a program *felicitous* if it embodies a happy hypothesis, i.e. if the falsehoods of the model serve to generate, via the program, exactly the truths of the model. Therefore, a felicitous model is one that, in this sense, leads to the whole truth and nothing but the truth. For example, the model which makes p true and q false is felicitous for the program $\{q \leftarrow -p\}$.

I wish to justify the rule of negation as failure in terms of the felicity of

models. The mere truth of a program in a model is not sufficient to guarantee that the statements inferred from the rule are true. Something more is required. I wish to suggest that the something more is that the model be felicitous, the falsehoods should give rise to exactly the truths.

The present suggestion may seem somewhat artificial; but in fact it has a very natural motivation in terms of the attempt to make sense of inductive definitions that need not be "monotonic" in the predicate or predicates that are being defined. As already explained, there is a natural notion of inductive definability for positive programs or definitions: the defined truths (or extensions) are the generated truths (or extensions). However, what if some clauses contain occurrences of negation within their body, so that a truth (or positive instance of a predicate) can be determined partly on the basis of falsehoods (or negative instances)? What if the definition of the predicate E for being even, for example, takes the form $\{Eo \leftarrow , Esx \leftarrow -Ex\}$. How then are we to conceive of the truths or the extensions of the predicates that are to be defined?

We might think of the generation process entirely in terms of truths or positive instances. The negative clauses are then in effect ignored, since their antecedents are never realized. This is not incorrect, but it provides no use for the negative clauses. The question of interest therefore is how such clauses might be used. How might falsehoods or negative instances enter into the generation process?

One possibility is to treat the statements which are not true at a given stage as false. (This corresponds to the standard treatment of non-monotonic definitions.) Thus at the start of the generation process, it is assumed that every statement is false. At the next stage, the clauses will generate certain truths, and so the remaining statements can be taken to be false. The process then continues in this way. It must still be determined what happens at the limit. The simplest option is to collect together all of the truths generated at previous stages, though there are various ways in which certain of these truths might be weeded out. (Similar remarks apply, both here and in the sequel, to the generation of positive and negative instances.)

This account does indeed make use of the negative clauses, but there appears to be something mistaken in the idea that we are entitled to assume that all statements are false at the beginning of the generation process. We are no more entitled to assume this than that all statements are true. In either case the assumption will usually be erroneous. It may get corrected in the course of the generation process, with falsehoods being supplanted by truths, but there is still no guarantee that the bad

effects of the negative facts which are subsequently removed will not remain.

Another possibility is to take a statement to be false only if there is no danger of showing it to be true. (This corresponds to a natural extension of the NF rule.) In effect, the program or definition is treated as implicitly containing clauses for the generation of falsehoods: a statement is taken to be false under the natural conditions which exclude its generation as a truth. The truths and falsehoods may then be generated simultaneously from the explicitly given clauses for the truths and the implicitly given clauses for the falsehoods.

This account is not excessive in its postulation of falsehoods, but it is unduly cautious. Even in the cases of ordinary inductive definition, we do not expect the falsehoods to be generable in the prescribed manner. They are simply taken, once the generation process for the truths is over, to be the remaining statements.

There are perhaps other accounts that might be suggested, intermediate between the two I have described. Perhaps we can *provisionally* assume that all statements are false at the beginning of the generation process. If this assumption gets us into trouble, we can revise it in an appropriate way to get out of the trouble and then start all over again. The generation process is continually revised until, it is hoped, a satisfactory process is reached.

However, the basic problem still remains. On the one hand, we want the final generation process to be sound: no falsehood can become a truth (or truth a falsehood). On the other hand, we want the generation process to be complete: every statement must eventually become true or false. Thus what a solution in general requires is that the falsehoods which we feed into the process should be the complement of the truths which we get out of the process.

To determine what it is that an extended inductive definition defines, we may attempt to anticipate in advance what the falsehoods (i.e. non-truths) will be. If we are right, then a set of truths will indeed have been defined. Such an anticipation of the falsehoods amounts to no more than a happy hypothesis. Thus the assumption that the model is felicitous provides one very reasonable account of what it is for a model to be inductively defined under an extended definition (or program).

Conversely, upon presupposing the present account of inductive definition, we can define what it is for a model to be felicitous. For we can say that such a model is one that makes all of the statements derivable from a program true, where the derivable statements are inductively defined as

those following from certain classical rules and a rule licensing $\neg A$ when A is not derivable. Indeed, it seems to me that many of the cases in which we are tempted to give a nonmonotonic inductive definition are ones that might reasonably be understood in terms of the present account. Examples abound in the literature on nonmonotonic reasoning, with derivability or some form of acceptance defined in terms of non-derivability or non-acceptance. But examples are also to be found elsewhere.

As is to be expected from the connection with inductive definition, the requirement that an intended model for a program be felicitous also has a formulation in terms of a closed-world assumption. Stated as a rule, the assumption takes the form:

CWA. If a positive statement is not derivable, then its negation can be inferred.

As already pointed out, the rule is not determinate until it is made clear what "derivable" means. Now one possibility is to explain derivability in terms of a pre-existing set of rules. This is what we did before, when the rules were those of classical logic or of Prolog. Another possibility is to include the rule CWA itself among the rules which determine what is derivable. Let us suppose that the other rules are specified as " R ". Then the CWA receives the self-referential formulation:

SRCWA. If A is not derivable by means of the rules SRCWA and R , then $\neg A$ can be inferred.

There are many possible choices for the auxiliary set of rules R , but the most relevant, for our purposes, is the one that enables the truths (or positive instances of a predicate) to be generated from the program clauses. Stated as a single rule, R is the result of combining applications of adjunction (infer $A \& B$ from A and B) with an application of *modus ponens* (infer B from $B \leftarrow A$ and A). It takes the form:

MPA. Infer B from A_1, \dots, A_n , given that $B \leftarrow A_1, \dots, A_n$ is a program clause.

Thus the rule MPA permits each of the program clauses to be translated into a rule of inference. Given that MPA constitutes the choice for the rule R , the self-referential version of the closed-world assumption becomes:

SRCWA. If A is not derivable by means of the rules SRCWA and MPA, then $\neg A$ can be inferred.

It should be noted that if a program is closed under the rules MPA and the self-referential CWA, then the resulting system will be consistent and complete. For if A is not derivable by means of the rules, then $\neg A$ is derivable by the CWA; while if A is derivable by means of the rules, then $\neg A$ is not derivable using the CWA and so not derivable at all. Completeness is guaranteed with any rule R in place of MPA; but consistency depends upon the fact that the rule MPA is only capable of yielding positive conclusions.

Given the self-referential character of the rule CWA, it is not clear what the rule means: for in order to know when the conditions for the application of the rule have been met, we need to know what conclusions the rule is capable of yielding. So how is the rule to be interpreted? A possible interpretation can be identified with the set of negative statements that can be inferred by means of the rule. A possible interpretation I is then permissible or coherent just in case it conforms to the condition that a negative statement $\neg A$ is in I iff A is not derivable using the rule MPA and the premises I . Thus a coherent interpretation of the rule is in effect a happy hypothesis: the adoption of the self-referential CWA amounts to the acceptance of a happy hypothesis.

It is in terms of this connection that we can extend the analogy between the two forms of inductive definition. As already noted, an ordinary inductive definition is converted into an implicit definition upon the addition of the regular extremal clause or CWA. In the same way, an extended inductive definition is converted into an implicit definition upon the addition of a self-referential extremal clause or CWA. Indeed, we can imagine that the extremal clause was, or should have been, self-referential all along. However, in application to positive definitions, the self-referential aspect was irrelevant; and it was only in application to extended definitions that this aspect came into its own.

All the same, there are some significant differences between the two kinds of definition. From the standpoint of the traditional theory of definition, the extended inductive definitions are in reality implicit definitions. The defining condition on the single predicate P , let us say, is that P should be inductively definable by such and such clauses from P' , where P' is the complement of P . Thus embedded in the implicit definition, if I can put it that way, is an inductive definition. (Of course, other relationships, besides complementation, could be required to obtain between the

defined predicates and the predicates from which they are inductively defined.)

Ordinary inductive definitions are entirely constructive in character: the truths (or positive instances of the defined predicates) can be built up by means of the clauses. The extended definitions have, by contrast, an essentially creative aspect: one must guess what the falsehoods (or negative instances) are before building up the truths (or positive instances); and it is only if the guess is subsequently confirmed that the construction is deemed correct. Of course, in certain special cases there may be a more direct account of what the truths or positive instances are; and it is a question of some technical interest to characterize such cases. However, in general, a guess will need to be made.

An ordinary inductive definition will uniquely determine the truths or the extensions of the predicates to be defined. For extended inductive definitions, uniqueness may fail, either because no set of truths or extensions is determined or because several are. This is already apparent from our examples of programs for which there are no happy hypotheses or for which there are several; for there will be the same number of coherent interpretations of the extremal clause as there are happy hypotheses. Again, it is a question of some technical interest to characterize the special cases in which there will be a unique hypothesis or interpretation for the program.

We may, of course, consider the intersection of the regularly defined sets of truths or falsehoods. This yields a 3-valued model which, like its parent interpretations, will validate the rule NF. However, such a super-valuational model is not satisfactory, either as an interpretation of the program or as an account of what is defined. Indeed, the model will not, in general, even make the clauses of the program or definition true.

Finally, extended inductive definitions are especially sensitive to the form of the auxiliary rule R . For ordinary definitions, it does not matter whether R is the simplest rule MPA or includes the full resources of classical logic; the class of derivable statements remains the same. For extended definitions, it does matter. Suppose, for example, that we are allowed to use a form of *modus tollens*, so that A can be inferred from $\neg B$ in the presence of the clause $B \leftarrow \neg A$. Then the program $\{q \leftarrow \neg p\}$ will admit another happy hypothesis or coherent interpretation of CWA; for given the hypothesis that q is the sole falsehood, it will follow by *modus tollens* that p is true. Indeed, given the full resources of classical logic, the program $\{q \leftarrow \neg p\}$ will admit the two happy hypotheses, p

false or q false, in contrast to the standard non-self-referential account of CWA, under which the program is rendered inconsistent.

Another kind of case is illustrated by the program $\{p \leftarrow \neg p\}$. With MPA as the sole auxiliary rule, there is no happy hypothesis, but upon the addition of the appropriate classical rule, the inference to p can be made and so “nothing false” becomes a happy hypothesis.

We see from these examples that within the field of extended definitions, there is not a single self-referential CWA which might be added to the body of a definition, but a whole range of them, which vary in accordance with the rule or rules that are taken to be auxiliary. I think that the case we have considered, in which MPA is the sole additional rule, is the simplest and most natural of them; but the others also deserve attention.

7. Self-referentiality vindicated

Let us now consider the merits of my own proposal as a justification of the rule of negation as failure. The proposal satisfies the purely formal requirement for a justification of the rule, viz. that the rule be sound. We will not present the full proof here, but will confine our attention to the critical case. We assume that the statement r fails and wish to show that its negation $\neg r$ is true, i.e. that the statement r itself is false in a felicitous model for the program. Suppose that r is not false. Then it is true. Thus, it is generated from some clause of the program, say $r \leftarrow p, \neg q$, where p and $\neg q$ are true. Now p cannot fail, since otherwise by the inductive hypothesis it is false, and not true; while $\neg q$ cannot fail, since otherwise q succeeds and so, by an appropriate inductive hypothesis, it is true, and not false. Therefore neither p nor $\neg q$ fail, contrary to the assumption that r fails.

The present proposal also solves the informal difficulties that were seen to confront the previous proposals. It was taken to be a great strength of Reiter’s account, in its application to positive programs, that it delivered the least fixed point, and a corresponding weakness of Clark’s account that it tolerated other fixed points. On this score, the regular and the self-referential CWA are on a par. For if the program is positive, then no falsehoods can be used in deriving truths and so the only felicitous model will be the least fixed point. (The equivalence between the regular and the self-referential CWA actually extends to a wider class of programs.)

It was taken to be a great weakness of Reiter's account that it rendered certain acceptable negative programs inconsistent and a corresponding strength of Clark's account that it did not. Certainly, the self-referential CWA does not suffer from the same difficulties on negative programs as the regular CWA. For as we have already seen, the program $\{q \leftarrow -p\}$ will be consistent under the one and yet inconsistent under the other.

Whether the self-referential CWA suffers from difficulties of its own is another matter. Every felicitous model (with MPA as the sole auxiliary rule) will be a fixed point; and so my account will inherit any of the excesses that might be thought to belong to Clark's account. In particular, the program $\{p \leftarrow -p\}$ will lack any felicitous models.

However, by adding further auxiliary rules we are able to exclude some of the consequences of Clark's account. We may adopt a rule, for example, that enables one in effect to replace a derived clause of the form $B \leftarrow -B, \dots$ with $B \leftarrow \dots$. The rule NF will still be sound; and yet such programs as $\{p \leftarrow -p\}$ will then enjoy felicitous models.

I myself am not sure that any modification to the self-referential CWA needs to be made. It seems to me that no ordinarily acceptable program is rendered inconsistent under the unmodified rule; and certainly it becomes much harder to check that the rule is coherent once other auxiliary rules are allowed. However, it should be of interest for those with different intuitions that these wider classes of programs can be accommodated within a self-referential stance.

The self-referential CWA also performs much better than the regular CWA on the issue of availability. It will be recalled that it was taken to be a defect of the regular CWA that the justification it provided was not in general available to the programmer; for he would have no ready access to which statements were derivable from the program. It is, by contrast, relatively easy to check that the conditions for the use of the self-referential rule are met, for there is no longer any need to determine which negative statements can be derived. We can simply take for granted the falsehoods in the language of the program and then see, on this basis, whether it is exactly the truths that can be derived. Consider, for example, the program $\{Eo \leftarrow, Esx \leftarrow -Ex\}$. Use \bar{n} for o preceded by n s 's. Then to check the conditions for the applicability of the self-referential CWA, we need only note that $E\bar{o}$ is derivable, that $E\bar{2n}$ is derivable for positive n since $E\bar{2n-1}$ is false, and that $E\bar{n}$ is not derivable for odd n since $E\bar{n-1}$ is not false.

It is also readily intelligible that the computational demands made on a

program should lead to the satisfaction of this requirement. Ideally, we would want a program to deliver all of the truths and all of the falsehoods, but this demand is unreasonable, since there is in general no satisfactory way to track the falsehoods. A more realistic demand therefore is one that takes the derivability of the falsehoods for granted (as given, if you like, by an outside authority) and then requires that all of the truths should be derivable by honest toil.

It therefore seems plausible that in those situations in which the rule NF is characteristically used, the self-referential CWA will hold. If this is correct, then a further question naturally arises, as we would like to know what are the consequences of the assumption for those situations; we would like to know what further rules, besides NF, will be valid under the assumption.

There is indeed a major new rule that is validated by the self-referential CWA; and it is one, moreover, that is highly relevant to the possibility of extending the inferential apparatus of Prolog. (I might add that, even with this new rule, Prolog will not enjoy supra-classical completeness with respect to the present justification.) Let us review the existing rule NF. This says that every failed statement can be denied. A statement fails if every path of an evaluation tree for the statement fails; and a path fails if it ends in failure, i.e. in a non-empty goal containing a negated atom that is successful (on the identity substitution) or an unnegated atom that is irreducible. Any failed path must, of course, be finite; and it is for this reason that the concept of failure relevant to the rule NF is called *finite failure*.

But what of the infinite paths? If success is termination in an empty goal, then it is natural to take an infinite path as constituting failure. We are therefore led to the following extension of the original rule NF. A path *fails* if it is either finite and fails in the usual way or if it is infinite; a statement *fails* (as before) if every path on an evaluation tree for the statement fails. The extended rule then says that a statement which fails in this general sense may be denied.

It is important, in this regard, to distinguish between the infinite evaluation induced by $\{p \leftarrow p\}$, on the one hand, and $\{p \leftarrow \neg p\}$, on the other. The first program leads to an infinite path. The second does not; it merely leads to the indefinite invocation of the same finite tree for the evaluation of $?p$. It is only the first kind of infinite evaluation, *within* a tree, that is germane to the application of the extended rule.

It was previously noted that the original rule NF can be formulated as a rule of inference within the context of a deductive system of the standard

sort. When an attempt is made to formulate the extended rule in the same way, it leads to the curious result that the derivations may be ill-founded; they may stretch indefinitely back. Consider, for example, the program $\{p \leftarrow p\}$. The original rule of inference corresponding to NF allows us to infer $\neg p$ from $\neg p$ within the context of this program. The second $\neg p$ can in the same manner be inferred from $\neg p$; and so on for any finite number of steps. Clearly, it is not thereby possible to arrive at a categorical assertion of $\neg p$. Suppose, however, that we took the derivation of $\neg p$ from $\neg p$ back through infinitely many steps. Then the rule of inference corresponding to the extension of NF would allow us to assert $\neg p$ on the basis of this infinite derivation. A circular argument would actually yield conclusions!

It should be intuitively clear that this new rule is validated by the self-referential CWA. For if a path of the evaluation tree for a statement is infinite, there cannot exist any means whereby the statement can be generated as a truth, even if all of the falsehoods are given. It is also reassuring that the assumption permits this extension. For one is inclined to think that there is no logical difference between finite and transfinite failure, that the one provides as good a basis for denying a statement as the other, and that it is only from considerations of implementation that the usual rule is confined to finite failure.

Once it is granted that the new rule is cogent in those situations in which the original rule is used, it is natural to seek ways of extending Prolog so that it may capture at least some of the content of the new rule. There is, of course, no possibility of conducting an infinite search through the evaluation tree in order to detect which of the paths are infinite. It must somehow be anticipated which of the paths are infinite so that a fail can be declared in advance of an actual traversal of the path. In some case it is clear how this is to be done; simple loops, as in the program $\{p \leftarrow p\}$, for example, can be blocked by a fail. The real question is whether there is a reasonably comprehensive and practicable procedure for anticipating the infinite paths.

This question would be significant even without a failure rule because time spent traversing an infinite path is wasted time; and so it would always be better to anticipate that a path is infinite, if one could, rather than proceed along it. But there is a difference between anticipating infinite paths for the purpose of avoiding them and for the purpose of fuelling a concept of failure. What I wish to suggest is that should procedures for anticipating infinite paths be devised, then they should be incorporated into the rule of negation as failure. The rule was conceived

in sin, with efficiency taking the place of validity; and its development can equally well be guided by sin.

8. The treatment of terms

We have so far engaged in the convenient fiction that the programs are constructed from simple sentence-letters. We must now consider the complications which arise from adding relation and function symbols to the language.

How is the resulting language to be interpreted? It is commonly assumed that the intended model for such a language is a *term* model: the domain is a set of ground terms, either from the language itself or an extension of the language; and the function symbols are so interpreted that each ground term will denote itself. However, it needs to be emphasized that the intended model will not, except in the rarest cases, be a term model: the domain will be of ordinary objects, such as numbers or people; and the function symbols will be for ordinary functions, such as addition or date of birth. There will, of course, be a term model corresponding to a "real" model; for we may take the terms to be those from the language of the real model, supplemented if need be with constants for its objects; and we may take a predicate P to be true of the terms t_1, \dots, t_n when the atomic sentence $Pt_1 \dots t_n$ is true in the real model. The resulting model will make the same atomic sentences true as the original intended model. However, it will still not, unless by accident, be the intended model.

Perhaps it is thought that the programmer will only have a term model in mind. But why should he mean anything different by the language than the rest of us? Why should he talk of numerals or people's names when the rest of us talk of numbers or people? It would be a strange departure from the ideals of logic programming if its languages could only talk about themselves. This is a departure, which as far as one can see, the programmer should neither want to make nor need to make.

Perhaps it is thought that, for the purposes of studying program languages, only term models need be considered; the differences between real models which correspond to the same term model can be ignored. Now I do not think this is so. Indeed, I think the differences can run rather deep. Even if it is so, it is something that should be shown rather than presupposed. The real models should at least be entertained if only for the purpose that they might subsequently be eliminated.

The pre-occupation with term models has no doubt arisen from an exclusive concern with the computational meaning of a program. It is as if the programmer becomes a solipsist when he sits at his console — only the world of the program and its actions exists; the rest of the world drops away. There is thus in this attitude a reversion to an old fashioned (and what one would have thought was a discredited) form of formalism.

The standard justifications of the rule NF all take the intended model for a program to be a term model or to have some of the distinctive features of a term model; and to that extent, they are suspect. The CWA account, at least in the formulation given by SHEPHERDSON (1984), requires the satisfaction (or addition) of Clark's Equality Axioms. These take the form of universal principles which one would expect to hold in all term models. They include, for example, the axioms $(x = t(x))$, where $t(x)$ is a term properly containing x ; and they have as a consequence that distinct ground terms denote distinct things. The account requires, in addition, the use of the Domain Closure Axiom, to the effect that every object is denoted by a term.

The completed domain account also requires the Equality Axioms (though not the Domain Closure Axiom). It requires, in addition, a general form of the biconditionals, one for each predicate of the language. So if the program is $\{Pa \leftarrow, Pfx \leftarrow Px\}$, the biconditional for P is:

$$\forall y(Py \leftrightarrow y = a \vee \exists x (y = fx \ \& \ Px)) .$$

These assumptions drastically curtail the scope of any possible justification of the rule NF. They make it impossible, for example, to justify the rule in those situations which are the norm rather than the exception, the situations in which two distinct terms will denote the same thing under the intended interpretation. It is presumably only because it has been assumed that the programmer has a term model in mind that these restrictive assumptions have been regarded with such equanimity.

But may this not be one of the cases in which only the properties of the term model need be considered? Let it be granted that the intended model deals with ordinary objects and functions. Still, associated with the real model will be a term model; and we can take it that the rule of negation as failure is justified for the real model when it is justified for the corresponding term model.

The trouble with this approach is that it places a much too stringent condition on the real models, ruling out many cases that one can and would want to consider. Thus the CWA requires the use of the Domain

Closure Axiom, which prevents any real model from containing undesigned individuals. Nor is this restriction harmless on consequences. For with it, $\forall xPx$ becomes a consequence of the program $\{Pa \leftarrow\}$.

One might, of course, enlarge the language of the real model with constants for its undesigned individuals, but the CWA would then fail in all but the most trivial cases. With the program $\{Pa \leftarrow\}$, for example, the statement Pb , where b is a constant for an undesigned individual, would not be a consequence of the program; and so the model would be forced to make Pb false. In general, the statement $A(b)$, where b is a constant for an undesigned individual, would only be a consequence of the program if $\forall xA(x)$ was; and so a model would either have to make every individual satisfy $A(x)$ or make every undesigned individual fail to satisfy $A(x)$.

The difficulty for the completed domain account may be illustrated with the program $\{Pa \leftarrow, Pfx \leftarrow\}$ and the model of two individuals i and j , in which a denotes i , P is true of i and j , and f denotes a function which takes i into i and j into j . Then it is perfectly reasonable to apply the rule NF to such a program (in fact, no statements will fail). However, the general biconditional governing the predicate P will be false in the term model corresponding to this model. For the biconditional takes the form:

$$\forall y (Py \leftrightarrow y = a \vee \exists x (y = fx)) .$$

The term model contains a new constant b for the individual j and has every ground atomic statement true. But instantiating y with b , we obtain:

$$Pb \leftrightarrow b = a \vee \exists x (b = fx)$$

which in the term model is false. In general, when some of the objects in the real model are undesigned, the presence of the new constants is likely to foul up the truth of the biconditionals in the term model.

It is in fact possible to tailor the closed-world and completed domain accounts much more closely to the functional structure of the real models. The restriction to term models and their like can then in a non-trivial way be removed. The approach has other advantages: for it serves the demands of mathematical generality; and it makes unproblematic the use of a standardly interpreted identity predicate within the clauses of a program.

Under the closed-world account, the Equality Axioms are not in fact

required. The Domain Closure Axiom is. Suppose, for example, that the program is $\{Qa \leftarrow Px\}$. Then Px fails; but to get $\forall x - Px$ from $-Pa$ we need the axiom $\forall x(x = a)$. In place of this axiom, one might use the rule which allows one to infer the identity-free statement $\forall xA(x)$ when all of its ground instances have been derived. Closure under the rule follows from the axiom. But one might accept the rule, if one thought, for example, that the designated individuals were representative of all individuals, without accepting the axiom. Still, this solution is not very satisfactory, either with or without an extension in the language of constants; for it either produces unwanted generalities or else treats all undesignated individuals alike.

The completed domain account is somewhat harder to deal with. We cannot simply drop the Equality Axioms since then the rule NF will not be vindicated. Under the program $\{Qa \leftarrow, Pb \leftarrow\}$, for example, the atom Pa fails. But to derive $-Pa$ from the biconditional for P , viz. $\forall x (Px \leftrightarrow x = b)$, it appears that we need the additional axiom $-(a = b)$. We could, of course, use a large number of biconditionals, one for each ground atom, in place of the general biconditionals. This might make the biconditionals themselves of infinite length; but it would not, in any case, be possible to account properly for the failure of unground atoms.

The correct approach is as follows. Given a program P and a model M for the language of P , let us extend L in the usual way with new constants for the undesignated individuals of M . Given a ground atomic statement A of the extended language L^+ , we say that another such statement B of L^+ is an *elaboration* of A if it is the result of replacing new constants c in A with ground terms t from L^+ with the same denotation as c in the model M . Thus an elaboration allows us to give more information about the newly denoted individuals.

Clark's account, when translated into a condition on the term model over L^+ corresponding to M , requires that every atomic truth of L^+ should be generable, by means of a clause of the program, from truths of L^+ . We require the much weaker condition that every atomic truth of L^+ should have an elaboration which is generable from truths. As long as we provide enough information about the undesignated individuals, each atomic truth is generable from truths.

The difference between the two accounts can be illustrated by our previous example of a program $\{Pa \leftarrow, Pfx \leftarrow\}$, whose model has two individuals i and j , with a denoting i , P a universal predicate, and f a symbol for an identity function. We introduce a new constant b for j . Now Pa and any ground atomic statement of the form Pft is generable (and

hence generable from truths). This leaves Pb ; but it has the elaboration Pfb , which is also generable. Thus the application of NF is justified, in contrast to the result for Clark's unmodified account.

A formulation in terms of real models can also be provided for the 3-valued version of Clark's account; but the details are somewhat more complicated and will not be considered here. However, we may note that the use of real models allows one to give a more natural completeness result for positive programs on either the 2- or the 3-valued approach. The usual completeness result is, in a certain sense, a fraud. For in order to account for the non-failure of Px in the program $\{Pfx \leftarrow Px\}$, we have to use a model that is not a term model. However, it is hard to see on what basis, besides the restriction to term models, the Equality Axioms could be accepted. It might be suggested that the standard term models somehow be expanded to include such infinite terms as $fff. . . a$; but then it would be implausible to retain the special axioms $(x = t(x))$ (not that it would be *impossible* to develop a theory of infinite terms that respected such axioms).

These difficulties disappear upon the admission of real models; since no *a priori* restriction is imposed upon the interpretation of the constants and function symbols. Indeed, it becomes much easier to find simple and non-artificial counter-models. For example, in the program above, we may let the model have two individuals i and j , with P true of one of them, say i , and with f denoting an identity function. The conditions for the justification of the rule NF are satisfied, since the truth of Pb (where b denotes j) has an elaboration Pfb which is generable from Pb ; and Px is not true when x takes the value i .

Finally, we may consider how my own account, in terms of the self-referential CWA, is able to accommodate real models. If we simply translate the real model into a term model, we obtain the very stringent condition that it is exactly the atomic truths of the extended language which can be generated from the atomic falsehoods. This means, for example, that if a predicate P is true of an undesigned individual j , then it must be true of every individual, since the generation of Pb , where b names j , cannot make use of any special properties of j .

Just as in the case of Clark's account, this condition can be replaced by a much weaker one. We need only require that no atomic falsehoods be generable and that every atomic truth have an elaboration which is generable. In the same way, it can be required, under the *regular* CWA, that every truth have an elaboration which is derivable. Thus the generation of an atomic truth concerning undesigned individuals need

not be completely generic; it can make use of as much information as can be packed into the terms by which reference to the undesigned individuals is made.

Acknowledgement

I should like to thank David Kaplan for helpful conversions on the topic of happy hypotheses and Stott Parker for pointing out infelicities in the original version of the paper. Since finishing the paper I learnt that some similar ideas have been independently proposed by GELFOND and LIFSCHITZ (1988); their stable models correspond to my felicitous models. These authors also point out the connection with the stable expansions of MOORE (1985).

References

- CLARK, K.L., 1978, *Negation as Failure*, in: H. Gallaire and J. Minker, eds. *Logic and Databases* (Plenum Press, New York), pp. 293–322.
- FITTING, M.C., 1986, *A Kripke-Kleene Semantics for Logic Programs*, *J. Assoc. Comp. Mach.* 29, pp. 93–114.
- GELFOND, M., and LIFSCHITZ, V., 1988, *The Stable Model Semantics for Logic Programming*, in: R. Kowalski and K. Bowen, ed., *Logic Programming: Proceedings of the Fifth International Conference and Symposium* (M.I.T. Press, Cambridge), pp. 1010–1080.
- KUNEN, K., 1987, *Negation in Logic Programming*, *J. Logic Programming* 4, pp. 289–308.
- LLOYD, J.W., 1984, *Foundations of Logic Programming* (Springer-Verlag, Berlin), 2nd ed., 1987.
- MOORE, R., 1985, *Semantical Considerations on Nonmonotonic Logic*, *Artificial Intelligence*, 25, pp. 75–94.
- REITER, R., 1978, *On Closed World Data-bases*, in: H. Gallaire and J. Minker, eds., *Logic and Data-Bases* (Plenum Press, New York), pp. 55–76.
- SHEPHERDSON, J.C., 1984, *Negation as Failure I*, *J. Logic Programming* 1, pp. 51–79.
- SHEPHERDSON, J.C., 1985, *Negation as Failure II*, *J. Logic Programming*, 2, pp. 185–202.
- SHEPHERDSON, J.C., 1988, *Negation in Logic Programming*, in: J. Minker, ed., *Foundations of Deductive Data-bases and Logic Programming* (Morgan Kaufmann, Los Altos).

This Page Intentionally Left Blank

FIRST-ORDER SPACETIME GEOMETRY

ROBERT GOLDBLATT

Mathematics Department, Victoria University, Wellington, New Zealand

1. Introduction

The theme of this study is the role of *orthogonality* as a primitive notion in metric affine geometry. To place this in perspective, recall that Hilbert's well-known foundation for Euclidean geometry can be regarded as based on the following two primitives:

- (1) the *linear betweenness* relation $B(xyz)$ ("y lies between x and z"), which provides a basis for *ordered affine geometry*: the geometry of linear subspaces and their translates in vector spaces over ordered fields.
- (2) the relation $xy \equiv zw$ of *congruence* of line segments ("the distance from x to y equals that from z to w"), which allows the introduction of *metric* notions (length, measurement of angles etc.).

Here we shall replace \equiv by the relation $xy \perp zw$, expressing "the line through x and y is orthogonal to that through z and w". In Euclidean geometry, this of course means that the lines xy and zw are *perpendicular*, but there are other geometries, including Minkowskian spacetime, in which \perp has quite different meanings, having both geometrical and physical significance. We shall set out complete and decidable axiomatisations of the first-order theories, in the language of B and \perp , for a number of such geometries.

2. Algebraic definition of \perp

A *metric vector space* (V, \cdot) over a field F consists of a vector space V over F with an *inner product* \cdot , i.e. a function $\cdot : V \times V \rightarrow F$ that is

- (1) bilinear: $x \cdot y$ is linear in each of x and y ; and
 (2) symmetric: $x \cdot y = y \cdot x$.

Vectors x and y are defined to be *orthogonal*, $x \perp y$, if $x \cdot y = 0$. Lines are then defined to be orthogonal if they have orthogonal direction vectors.

If V has dimension n , with basis $\{v_1, \dots, v_n\}$ relative to which vectors x and y have coordinates $\vec{x} = (x_1, \dots, x_n)$, $\vec{y} = (y_1, \dots, y_n)$, then bilinearity of \cdot entails that

$$x \cdot y = \vec{x}G\vec{y}^T$$

where G is the $n \times n$ -matrix whose i - j -th entry is $v_i \cdot v_j$. G is a symmetric matrix, by the symmetry of the inner product \cdot .

This construction shows that inner products on an n -dimensional vector space V , and hence orthogonality relations on the set of lines of V , are determined by symmetric $n \times n$ matrices over the scalar field F . By standard techniques of linear algebra, V can be coordinatised in such a way that G is diagonal. Moreover, if F is a *quadratic* field, i.e. each element or its opposite has a square root in F , the coordinatisation can be arranged in such a way that each diagonal entry is 1, -1 , or 0.

3. Planes over \mathbb{R}

In the next four sections we describe the geometries to be axiomatised. In the case that $V = \mathbb{R}^2$, there are three significant inner products, as follows.

3.1. Euclidean plane

The identity matrix $G = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ gives the inner product $x \cdot y = x_1y_1 + x_2y_2$, whose associated orthogonality relation is the relation of perpendicularity.

3.2. Lorentz plane

Here $G = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and $x \cdot y = x_1y_1 - x_2y_2$. This is two-dimensional spacetime, used to chart the history of an object moving in one spatial dimension, represented by the horizontal axis. The first notable feature of \perp here is the presence of *self-orthogonal* ($L \perp L$) lines, namely those of slope ± 1 . Such lines, also called *null*, or *lightlike* in the case of spacetime,

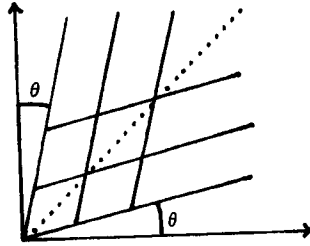


Fig. 1.

are the worldlines of particles moving at the speed of light. Lines having slope between 1 and -1 are *spacelike*, while the remaining lines are *timelike*. Spacelike lines are characterised by the condition $x \cdot x > 0$, while timelike lines have $x \cdot x < 0$.

If two non-null lines in the Lorentz plane are orthogonal, then one will be timelike, the other spacelike, and their slopes will be mutually inverse. The timelike line will be the worldline of an observer moving at constant speed relative to the “stationary” frame of reference given by the x and y axes, while the spacelike line consists of points that are regarded as *simultaneous* by that observer. This physical interpretation of orthogonality leads to the picture of Fig. 1, in which the spacelike lines of slope $\tan \theta$ are lines of simultaneity for an observer whose timelike worldline has slope $1/\tan \theta$. The two frames of reference have the same lightlike (dashed) lines, as dictated by the principle of the absoluteness of the speed of light.

Notice that when two lines are orthogonal in the Lorentz plane, the product of their slopes is 1, whereas in the Euclidean plane the product of slopes of perpendicular lines is -1 . The existence of such a *constant of orthogonality* is a key to the structure of metric planes in general (cf. §8).

3.3. The Robb plane

The Robb plane is defined by the inner product $x \cdot y = x_1 y_1$, generated by the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ on \mathbb{R}^2 . Under the resulting orthogonality relation, all vertical lines are *singular*, i.e. orthogonal to all lines in the plane, including themselves. All other lines are orthogonal only to the vertical singular lines.

The Robb plane is named after A. A. Robb, whose book *A Theory of Time and Space* (1914) gave the first detailed account of its structure. Copies of this two-dimensional geometry (called *optical planes* by Robb)

arise in three-dimensional spacetime as tangent planes to the lightcone (cf. Fig. 2).

4. Non-singular three-spaces

A metric vector space is *non-singular* if it has no lines that are singular, i.e. orthogonal to all lines (this is equivalent to the matrix G being non-singular, i.e. invertible).

4.1. Euclidean three-space

Euclidean three-space, defined by the 3×3 identity matrix, is a non-singular metric vector space over \mathbb{R}^3 . The only other one is the *three-dimensional Minkowskian spacetime*, with matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

In this case the null lines through the origin form a three-dimensional cone, the lightcone, as depicted in Fig. 2. These null lines are the worldlines of particles moving out from the origin in all directions on a two-dimensional surface represented by the horizontal plane through the origin. Any line interior to the lightcone is timelike ($x \cdot x < 0$) and is the worldline of an observer moving at constant speed. The lines through the origin that are orthogonal to this timelike line are themselves all spacelike ($x \cdot x > 0$), and form a *plane of simultaneity* for that observer, this plane having inverse slope to that of the worldline. Such a plane of simultaneity is isomorphic to the Euclidean plane. On the other hand, the set of lines through the origin that are orthogonal to a given spacelike line form a plane isomorphic to the Lorentz plane, cutting the lightcone along two null lines. Finally, the lines orthogonal to a given lightlike line form a

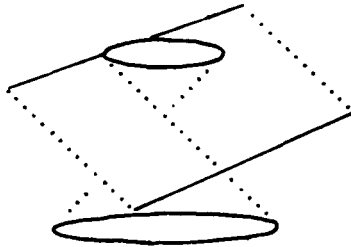


Fig. 2.

tangent plane to the lightcone: an isomorphic copy of the Robb plane, as already indicated.

5. Singular three-spaces

5.1. The singular matrix

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

induces the orthogonality structure on \mathbb{R}^3 illustrated in Fig. 3. The null (self-orthogonal) lines consist of all those in the y - z -plane. These lines are singular in \mathbb{R}^3 , and all other lines are orthogonal only to them. Thus any other plane through the origin is isomorphic to the Robb plane.

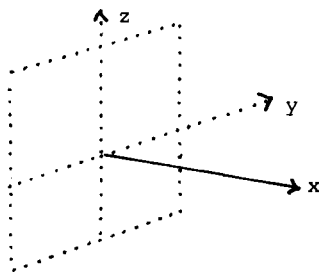


Fig. 3.

5.2. The matrix

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

gives the geometry of Fig. 4, in which the null lines are those of the two planes $y = x$ and $y = -x$. Each is a null plane, i.e. any two lines within the plane are orthogonal, but the z -axis is the only line through the origin that is singular in \mathbb{R}^3 . Any plane containing the z -axis, other than the two null planes, is isomorphic to the Robb plane, while the remaining planes through the origin are isomorphic to the Lorentz plane.

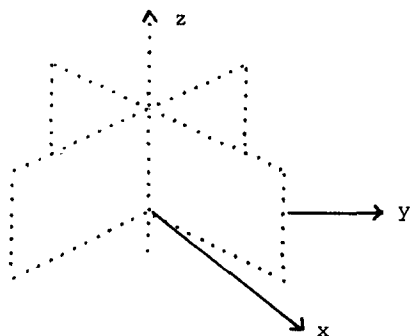


Fig. 4.

5.3. The matrix

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

generates *three-dimensional Robb space*, in which the vertical axis is singular in \mathbb{R}^3 , and is the only null line through the origin. Planes containing this axis are isomorphic to the Robb plane, while all other planes through the origin are isomorphic to the Euclidean plane. This structure, called an *optical threefold* by Robb, is manifest by the tangent spaces to the lightcone in four-dimensional spacetime.

6. Non-singular four-spaces

In addition to Euclidean four-space (whose inner product is given as usual by the identity matrix), there are two non-singular metric geometries over \mathbb{R}^4 .

Minkowskian spacetime is generated by the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

The physical interpretation is as before. Self-orthogonal lines are worldlines of particles moving at light-speed. The lines through the origin

orthogonal to a given line L form a *three-space*. If L is timelike, this is a copy of Euclidean three-space and is a *space of simultaneity* for an observer whose worldline is L . If L is spacelike, the orthogonal three-space is isomorphic to three-dimensional spacetime, while if L is null it is isomorphic to Robb three-space and is tangential to the lightcone.

Artinian four-space has its inner product determined by the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Here the “lightcone” has equation

$$x_1^2 + x_2^2 - x_3^2 - x_4^2 = 0$$

and its cross-section through the three-space $x_4 = 1$ is the surface (hyperboloid) $x_1^2 + x_2^2 - x_3^2 = 1$ depicted in Fig. 5. Each point on this surface represents a null line through the origin in Artinian four-space. In fact the surface is ruled by two families of straight lines, each family consisting of mutually skew lines. Through each point p on the hyperboloid there passes one line from each family $-L$ and M as illustrated. Each of these lines is the cross-section of a null plane in Artinian four-space, while the plane containing L and M in the three-space $x_4 = 1$ is itself the cross-section of a singular three-space of the type described in 5.2, with p representing the singular line through the origin in that three-space.

Note that Minkowskian spacetime has no null planes, and hence no singular subspaces of the type described in 5.2. Neither it nor Artinian four-space has singular subspaces of the 5.1 type.

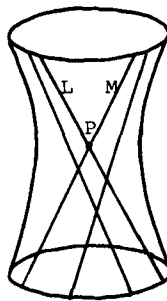


Fig. 5.

7. Ordered affine geometry

If a field F has an ordering \leq , then in any vector space V over F , the linear betweenness relation $B(xyz)$ can be defined by the condition

$$\exists \lambda (0 \leq \lambda \leq 1 \wedge y = (1 - \lambda)x + \lambda z). \quad (\dagger)$$

Now in the first-order language L_B of a ternary relation symbol B , there is, for each n , an explicit set OA_n of sentences such that for any L_B -structure \mathcal{A} ,

$$\mathcal{A} \models OA_n \quad \text{iff} \quad \mathcal{A} \cong (F^n, B) \text{ for some real-closed ordered field } F.$$

Recall that an ordered field is *real-closed* if its positive elements have square roots, and its odd degree polynomials have zeros. Such a field is elementarily equivalent, in the language L_{OF} of ordered fields, to the real number field \mathbb{R} . Thus the first-order theory of real-closed ordered fields is the same as the theory of the field \mathbb{R} , and this theory is complete, recursively axiomatisable, and hence decidable.

Now the definition (\dagger) allows, for each n , a translation of L_B -formulae ϕ into L_{OF} -formulae ϕ^+ , such that each L_B -sentence σ satisfies

$$(F^n, B) \models \sigma \quad \text{iff} \quad F \models \sigma^+.$$

First each variable x is assigned a list x_1, \dots, x_n of distinct variables. Then for atomic formulae, $(x = y)^+$ is

$$x_1 = y_1 \wedge \dots \wedge x_n = y_n$$

while $B(xyz)^+$ is

$$\exists \lambda (0 \leq \lambda \wedge \lambda \leq 1 \wedge \phi_1 \wedge \dots \wedge \phi_n),$$

where ϕ_i is

$$y_i = (1 - \lambda) \cdot x_i + \lambda \cdot z_i.$$

The translation is then completed by allowing it to commute with the propositional connectives, and putting

$$(\forall x\phi)^+ = \forall x_1 \cdots \forall x_n (\phi^+).$$

By the above observations about real-closed fields, we have

$$F \models \sigma^+ \quad \text{iff} \quad \mathbb{R} \models \sigma^+,$$

whenever F is real-closed, and hence in this case that

$$(F^n, B) \models \sigma \quad \text{iff} \quad (\mathbb{R}^n, B) \models \sigma.$$

Thus all models of OA_n are elementarily equivalent, and so OA_n is a complete and decidable theory.

As is well known, the results briefly reviewed in this section are due to Tarski.

8. Planar axioms for \perp

In the language L_B one can readily define formulae $Col(xyz)$ and $Cop(xyzw)$ expressing “ x , y , and z are colinear” and “ x , y , z , and w are coplanar”, respectively. The first formula asserts that one of x , y , and z lies between the other two, while the second uses Col to state that w lies on a line that passes through one vertex of triangle xyz and meets the opposite side. With these definitions we can specify the basic planar axioms for \perp as follows.

$$O1. \quad xy \perp zw \rightarrow zw \perp xy$$

$$O2. \quad [Cop(xyzw) \rightarrow xy \perp zw] \vee$$

$$\exists t [Cop(xyzt) \wedge \forall u (Cop(xyzu) \rightarrow (xy \perp zu \leftrightarrow Col(tzu)))]$$

$$O3. \quad xy \perp zw \wedge xz \perp yw \rightarrow xw \perp yz$$

O1 asserts the symmetry of the relation \perp on lines. O2 may be rendered as: in the plane of x , y , and z , either xy is singular, or there is exactly one line through z orthogonal to xy . This embodies the procedure, familiar in Euclidean geometry, of constructing an *altitude* to a given line through a given point. Here we require the given line xy to be non-singular in the plane of x , y , and z , and caution that if xy is a *self-orthogonal* non-

singular line, as can occur in the Lorentz plane, the altitude to xy through z will actually be *parallel* to xy (this is illustrated in the last two diagrams in Fig. 6).

Axiom O3 is manifest in a number of ways, illustrated in Fig. 6, depending on which lines involved (if any) are self-orthogonal (dashed). The first diagram of Fig. 6 indicates that the axiom implies the familiar Euclidean property that the altitudes of a triangle are concurrent, and indeed this property is equivalent to O3 in any affine plane that satisfies O1, and O2, and has no singular lines. On the other hand, the last diagram of Fig. 6 shows that O3 also encapsulates a property that is distinctive of the geometry of the Lorentz plane, namely that any parallelogram of self-orthogonal lines has orthogonal diagonals.

The properties of \perp expressed by O1–O3 can be asserted for the lines of the affine plane over any field F , and suffice to allow the construction of an inner product over F that coordinatises \perp (a significant side-point is that in a Desarguesian plane, these axioms imply Pappus's Theorem, which forces the coordinatising algebra to be a field rather than just a division ring).

For a plane with a singular line, a direct analysis of \perp , using O1 and O2, shows that it has the structure of the Robb plane: there is a parallelism class of singular lines, with all other lines orthogonal only to these. For nonsingular planes, there are two methods of coordinatisation available, as follows.

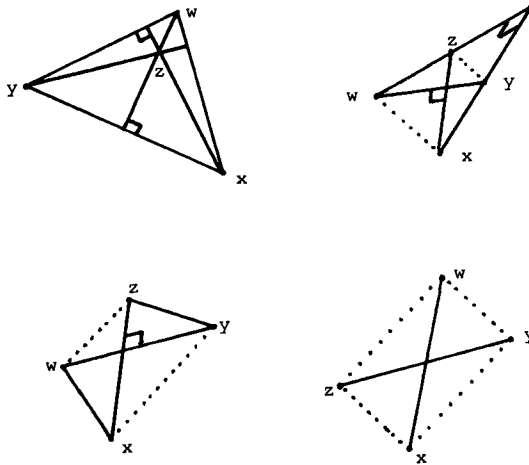


Fig. 6.

Method 1 fixes a point x in the plane, and uses O_2 to define a transformation on the pencil of lines through x , assigning to each such line its altitude through x . This becomes a mapping along the line at infinity in the projective completion of the affine plane, which is of period two by O_1 , and a projectivity by O_3 . There is a standard theory of matrix representation of such mappings (involutions) which yields a symmetric matrix giving the desired inner product. This method is a reversal of the classical procedure of defining perpendicularity in the Euclidean plane by using an involution on a line (at infinity) in the real projective plane.

Method 2, originating with Reinhold Baer, demonstrates that there exists a $k \in F$ such that the slopes of any line and any of its altitudes have product k (cf. the remarks on constants of orthogonality at the end of §3.2). The matrix $\begin{pmatrix} -k & 0 \\ 0 & 1 \end{pmatrix}$ then generates the coordinatising inner product. Analysis of the constant k gives a good deal of information about the orthogonality relation, e.g. that it has null lines iff \sqrt{k} exists in F , and allows one to conclude that over a quadratic field (in which each element or its opposite has a square root), there are up to isomorphism only two non-singular metric affine planes, one with no null lines ($k = -1$), and one with two null lines through each point ($k = 1$). When $F = \mathbb{R}$, these are of course the Euclidean and Lorentz planes respectively.

9. The three real planes

Let MOA_2 be the set OA_2 of axioms for ordered affine planes over real-closed fields, together with the orthogonality axioms O_1 – O_3 . Up to elementary equivalence, there are three models of MOA_2 : the Euclidean, Lorentz, and Robb planes. The theory MOA_2 has three complete and decidable extensions, namely the theories of these three models. Each theory is axiomatised by adjoining a single axiom to MOA_2 , as follows.

For the Euclidean plane, the axiom is

$$\forall x \forall y \neg \text{null}(x, y),$$

where $\text{null}(x, y)$ is the formula

$$(x \neq y) \wedge (xy \perp xy).$$

For the Robb plane, the axiom is

$$\exists x \exists y \text{sing}(xy) \wedge \exists z \exists w \neg \text{sing}(zw),$$

where $\text{sing}(xy)$ is the formula

$$(x \neq y) \wedge \forall z (xy \perp xz).$$

Finally, the axiom for the Lorentz plane asserts that there is a non-singular null line:

$$\exists x \exists y (\text{null}(xy) \wedge \neg \text{sing}(xy)).$$

10. Metric space axioms

For spaces of dimension three or more we need two further axioms for orthogonality.

$$\text{O4.} \quad [xy \perp yz \wedge xy \perp yw \wedge \text{Cop}(yzwu)] \rightarrow xy \perp yu$$

$$\text{O5.} \quad xy \perp zw \wedge zw \parallel uv \rightarrow xy \perp uv$$

O4 asserts that if line xy is orthogonal to lines yz and yw , then it is orthogonal to all lines through y in the plane containing yz and yw . Denoting by MOA_n the theory $OA_n + \text{O1-O5}$, we have that MOA_n is the first-order theory of metric affine spaces (F^n, \perp, B) given by n -dimensional inner product spaces (F^n, \cdot) over real-closed fields F .

11. The five real three-spaces

MOA_3 has five complete and decidable extensions. Three of these contain the sentence $\exists x \exists y \text{sing}(xy)$, asserting the existence of a singular line, and are the first-order theories of the three singular geometries over \mathbb{R}^3 described in §5. Each has a single characteristic axiom over MOA_3 .

$$(1) \quad \exists xyz [\neg \text{Col}(xyz) \wedge \text{sing}(xy) \wedge \text{sing}(xz)]$$

This asserts the existence of two intersecting singular lines. The plane containing two such lines is null and consists entirely of lines that are singular in the ambient three-space, as for the y - z -plane in Fig. 3.

$$(2) \exists xyz[\neg Col(xyz) \wedge sing(xy) \wedge \neg sing(xz) \wedge null(xz)]$$

This axiom states that there exists a singular line (such as the z -axis in Fig. 4) intersected by a non-singular null line.

$$(3) \exists xy[sing(xy) \wedge \forall z(\neg Col(xyz) \rightarrow \neg null(xz))]$$

Axiom (3) asserts that there is a singular line intersected only by non-null lines, and gives an axiomatisation of the three-dimensional Robb space described in §5.3.

The proof that these singular geometries are axiomatised as claimed involves a direct analysis of their orthogonality relations, together with the coordinatisation results for metric planes described earlier.

The remaining complete extensions of MOA_3 are the theory of Euclidean three-space, with characteristic axiom

$$\forall x \forall y \neg null(xy),$$

and that of the three-dimensional Minkowskian spacetime, with axiom

$$\forall x \forall y \neg sing(xy) \wedge \exists x \exists y (null(xy)).$$

The coordinatisation method for these last two spaces is described in the next section.

12. Axioms for spacetime

$MOA_4 + \forall x \forall y \neg sing(xy)$ is the theory of non-singular metric affine four-spaces over real-closed fields, and has, up to elementary equivalence, the three models described in §6. The characteristic axioms for the theories of these models is as follows.

For Euclidean four-space: $\forall x \forall y \neg null(xy)$.

For Artinian four-space:

$$\exists xyz(\neg Col(xyz) \wedge null(xy) \wedge null(xz) \wedge xy \perp xz),$$

which states that there is a pair of intersecting orthogonal null lines. The plane containing such a pair must be null, and Artinian four-space is the only non-singular four-space over \mathbb{R} containing null planes.

Finally, for Minkowskian spacetime, we negate the two axioms just

given, and assert that there exist null lines, but no two intersecting null lines can be orthogonal. This is expressed by the first-order sentence

$$\exists xy(\text{null}(xy)) \wedge \forall xyz(\neg \text{Col}(xyz) \wedge \text{null}(xy) \wedge \text{null}(xz) \rightarrow \neg xy \perp xz).$$

The coordinatisation procedure for a non-singular model of MOA_n ($n \geq 3$) generalises the first method for planes described in §8. It consists in fixing a point x , and analysing the bundle of affine lines, planes, three-spaces etc. that pass through x . These are the points, lines, planes etc. of the *projective* $(n - 1)$ -space over x . If L is an affine line through x , then the set L^\perp of all lines through x orthogonal to L is proven to be a hyperplane, i.e. an affine $(n - 1)$ -space, and the correspondence $f: L \rightarrow L^\perp$ is a *polarity*, a projective transformation which maps projective points to lines in the case $n = 3$, and points to planes when $n = 4$, whose restriction to any projective line is a finite sequence of projections and sections. The classical theory of matrix representation of polarities yields a diagonal matrix which defines an inner product representing \perp .

In proving that the stated axioms for spacetime, say, do generate a complete theory, the translation ϕ^+ of L_B -formulae into the language of real-closed fields is extended to formulae involving \perp by using the definition of the Minkowskian inner product to specify $(xy \perp zw)^+$ as

$$u_1 \cdot v_1 + u_2 \cdot v_2 + u_3 \cdot v_3 - u_4 \cdot v_4 = 0,$$

where u_i is $(x_i - y_i)$ and v_i is $(z_i - w_i)$.

The other geometries are treated in like fashion, translating $xy \perp zw$ in each case by the coordinate-wise definition of the relevant inner product.

A detailed account of the geometrical theory needed for these first-order axiomatisations appears in GOLDBLATT (1987).

Reference

- GOLDBLATT, R., 1987, *Orthogonality and Spacetime Geometry* (Springer-Verlag, New York), x + 190 pp.

6
General Methodology
of Science

This Page Intentionally Left Blank

STRONG AND WEAK METHODS

VOJTECH FILKORN

Slovak Academy of Sciences, Institute of Philosophy, Bratislava, CSSR

A method, in its broadest sense, is the way by which we follow the way, i.e. the character or the behaviour, of an examined object or domain. This means that a method involves two ways which, though different, are internally contingent. In order to know the way of an object we have to trace and follow it in a certain way. Therefore, the way of an object (*hodos*) determines the way of its subsequent inquiry (*met-hodos*). This means that the methods of investigation of an object or a domain will be determined by the nature of the object or the domain. From this point of view, a method is a reflection of the movement and of the substance of the domain under examination and so it has an ontologically objective character. Otherwise our way would not correspond to the way of a given domain and our whole “journey” would be not only useless, but also devious, wrong, and in that sense false.

On the other hand, however, we know that we are not capable of rational intuition which would give us a direct insight into the substance, into the internal way of an object or a domain, and so the inner part, the basis of the whole object from which its complete temporal concreteness is developed, is not immediately discerned. In fact, almost the opposite is the case—the substance of an object is revealed gradually only by its manifestations, reactions, etc. From this standpoint, our way, i.e. a method, is not identical with the way of the development of an object itself. So a method is of a relatively independent *gnoseological character*.

As we have no rational intuition, we are consequently historical creatures and the objective reality is not open to us in all its depth, complexity and width. We take hold of it only gradually, and step by step we reveal its general aspects and relations. Being discovered also in other domains, these aspects and relations are ascribed a universal *category* character. This character also later expresses our attitude to reality and

determines the ways of comprehension and research of other domains. It determines the framework of our insight into reality and our reactions to it. Thus we speak of a categorial ontological picture of the world.

History of science, when treated from this point of view, is the history of the gradual establishment and application of ontological categories which later either consciously or unconsciously become indicators of new trends, styles and ways of further research. In this way the ontological categories are transformed in our knowledge into methodological ones. Therefore, if some domain is to be successfully examined in a certain period, one must know all the previous ontological categories, and simultaneously pay attention to the eventuality that it may be necessary to introduce new views on reality which reflect its new aspects (in case some new categories appear). Thus, the unity of categories, by which a general categorial picture of the world is expressed, becomes an adequate methodological tool for its investigation. If some categories are either disregarded or not used, i.e. if they are not "changed" into methodological categories, we are led to a simplified or even to a false picture of reality.

We can conclude that the way of human cognition is determined by three types of categories (by three types of regularities).

A method, *as a progress towards a goal*, is submitted to the laws of this progress as such; this is expressed by the *gnoseological categories* such as truth, reflexion, concept, hypothesis, theories, language, prediction, explication, temporality of our knowledge, etc.

A method, as a reflection of the categorial ontological world, is determined by the *ontological categories*. This determines the general subject matter of our procedure. The ontological categories that may be listed here, are quality, quantity, relation, unity, movement, determination, phenomenon and substance, contradiction and so on.

The method itself consists of the orientation ofgnoseological categories on the ontological categories in the research of reality. Therefore the methodological categories are the forms of the ontological content ingnoseological categories. Ontological andgnoseological categories are connected through *methodological* ones as for instance: analysis, synthesis, the formation of concepts, hypotheses, and theories and their development, deduction, induction, etc. Methodological categories are a reflection of ontological categories in a double sense, which will be explained by the following example. An ontological category of unity in diversity is reflected in a method by synthesis and analysis. Similarly, the category of phenomenon and substance is reflected in a method by the

methodological categories of justification and explication but also by the method of inquiry from phenomenon to substance and the method of explanation from substance to phenomenon. A category of the unity of an individual—the particular—and the general is reflected by the categories of induction and deduction, etc.

However, individual methodological categories as reflections are of various forms, in accordance with the application of individual ontological categories. Thus, for example, the categories of quality, quantity, relation and contradiction, determine the subject matter of analysis and synthesis, so that a qualitative (classificatory), quantitative, relational and dialectical analysis and synthesis can be distinguished. Similarly, we can speak of classificatory. . . dialectic induction, justification and explication.

It is hardly possible to deal with all these problems here. Only one of the gnoseological categories will be examined—the way of the human cognition and activity itself and its structure.

Each human activity (and a cognitive process is also a sort of human activity) proceeds from some initial point toward a conscious or a less conscious aim; it is an activity, a process, a performance of successively correlated steps aimed at reaching an end or at attaining it gradually. The succession of steps and their successive correlation constitutes a *sequence*. As this activity is directed by precise effort to reach an aim, i.e. to know a certain domain, to solve a certain task, to construct a certain object, it must not be arbitrary or undetermined. Therefore, it can be realized only by means of certain steps successively correlated in a sequence. This suggests that the way towards an aim has to be a system. Thus a method *Me* is a dynamic system, i.e. it is a sequence *Po* of operations σ on the universe *M* of the applicability of the method, by means of which an aim (output) *V* is gained from a starting point (input, basis *Z*). Therefore a method can be characterized as

$$Me = \langle M, D, Z, V; \sigma, Po \rangle, \quad (1)$$

where $Z \subset D \subset M$. If a method is considered as a process, i.e. a determination of an aim from a starting point, it can be said that

$$V = Me(Z) \quad (2)$$

Thus *V* is a value of a global methodological function *Me* with an argument *Z*. The second sentence, (2), is valid only for strong methods. We explain the definition of the method (1) by characterising the meaning of *M*, . . . , *Po*.

The universe M of the applicability of the method consists of all domains D of researched or constructed reality, and of the domains of problems solved by the given method.

The heterogeneity of the initial points Z in a method corresponds to that of the aims of cognition and activity. It corresponds to the heterogeneity of the ways that can lead us to an aim. Generally, we can say, the initial point Z consists of what is (from the standpoint of the aim) given, known; usually it means a subdomain of a domain, or a phenomenal aspect of the whole domain. The aim is to understand the whole domain or, eventually, its full depth.

If the initial point is represented by the whole domain, its extension as well as its content will be examined. We assume that the extension as well as the content can be known, partially known, or unknown. Thus nine possible methodological situations are created. Each of these situations implies appropriate forms of procedures.

Now we explain the operation signs σ and sequences Po in (1). Operations are steps of procedure and elements of sequences. Single-valued operations ${}^1\sigma$ are the operations in which only one element corresponds to one or n elements; these operations are the strongest ones.

The subject matter of the operations is determined by the character of the domain under consideration, while each operation in a method is a step forward. The initial point of a step is determined by the independent variables, and its endpoint by dependent variables. In our activity, not all the steps are so simple. In order to give a true picture of even more complicated situations occurring in methods, we have to introduce a generalized notion of an operation, a *multi-operation*. A multi-operation is a multi-valued correspondence by which either zero or more elements correspond to an n -tuple of elements; an n -valued operation will be called ${}^n\sigma$.

Multi-valued operations occur in human activity as well as in scientific work quite frequently. They express our ignorance, uncertainty or undecidability. Multi-operations are used in those cases in which we do not know how to create on the basis of our knowledge of X_i ($i = 1, 2, \dots, m$) a single-valued picture H of this domain. Therefore we create alternative pictures (hypotheses) H_i ($i = 2, 3, \dots, n$) as the basis for future research. If we have two alternative pictures H_1, H_2 , we use double-valued operations ${}^2\sigma$

$${}^2\sigma(Xi) = H_1, H_2 \quad (3)$$

or

$${}^2\sigma(Xi) = H_1 \vee H_2 \quad (4)$$

where (4) implies two sequences Po , by which consequences are derived from H_1 and H_2 until one of these hypotheses is falsified. By this ${}^2\sigma$ changes into ${}^1\sigma$ and our inquiry becomes single-valued.

Null operations ${}^0\sigma$ occur frequently not only in science but in all kinds of human activity. ${}^0\sigma(x, y) = \emptyset$. In a method, a null operation is an operation in which no element corresponds to a value of the independent variables. This operation does not lead to an aim, but to a deadlock. It is a kind of methodological blind alley, which can, however, be useful, because it can be used for making procedures single-valued. Many blind alleys in research can be encountered in the history of science.

A sequence of operations Po can be understood as a correspondence of an operation to operations, i.e. as an ordered set of operations. Sequences can be single-ramified, multi-ramified, single-valued, multi-valued, eventually empty, linear or cyclic. Single-ramified and single-valued sequences are relatively simple. Multi-ramified sequences will be more complicated.

Multi-ramified sequences (trees) are chains of single-ramified sequences that make it possible for at least the last step of the sequences to join up all the branches in order to attain a single aim. A multi-ramified sequence is a sequence PPo on single-ramified sequences, so that $PPo = \sigma(Po_1, Po_2, \dots, Po_m)$. A multi-ramified sequence is therefore more complicated than single-ramified sequences. It is more suitable for the research of complicated domains, for the solution of complicated problems as well as for attaining complicated aims. This results from the fact that a method reflects the examined reality. The single branches of a tree can represent the research of the individual parts or aspects of a domain or the research of separate cases; and the tree represents their gradual joining into a unit or their eventual generalization. Trees can also express a gradual construction of individual parts and their gradual joining into complete technical devices.

The sequences in methods can be single-valued or multi-valued. Depending on whether the operations on the elements and the sequences are single-valued or multi-valued, we obtain many methods. All these methods must be designed in such a way as to allow the consequences to be single-valued in order to determine one aim.

Relatively complicated situations in research or in design have to be expressed by means of multi-valued sequences. Multi-sequences need not

be only an expression of our doubts but also of our effort to look for an optimal, e.g. the shortest way to an aim, of our endeavour based on the versatility of our view of the investigated domain, and so on. Multi-operations diverge sequences which are successively correlated. More complicated methods are obtained if we have more branches with multi-valued operations, but with single-valued sequences prevailing, immediately at the beginning. An even more complicated situation will occur if the sequences are also multi-valued. They lead to complete trees diverging from the beginning, i.e. to *furcations* which, if a method is to lead to an aim, must gradually become single-valued. We know from the history of science that it occasionally takes a considerable time to reach complete single-valuedness (as with the theory of light).

All the sequences mentioned above were basically linear. All linear methods presuppose a precisely determined initial point such that there is no need to return to it at a later stage. Such cases are not found in the empirical sciences. We often encounter a situation, and this is related to the lack of rational intuition, i.e. to the relative character of our cognition, that at the beginning of a methodological procedure the initial point is neither deduced nor completely substantiated. In that case, the initial point is only conditionally accepted and in that sense provisory and preliminary. Therefore, it can be substantiated only by the results of a methodological process. Thus a proceeding process, a gradual determination of a domain or a solution of a problem, is constantly accompanied by a retrospective, backward look, oriented towards the initial point. These two processes are methodologically identical. Therefore the steps forward are simultaneously regressive. Thus a progress from an initial point is simultaneously regressive, and so, as Hegel maintains, this method moves in a circle and gains a *cyclic* character. The results of a process and their constant confrontation with reality make us modify, accommodate, even change our initial suppositions, and thus we are even forced to modify the whole subsequent process; so it seems as if we were to start always from the very beginning. This, of course, does not mean that it is always the same beginning, because in that case no progress could exist in science and the repetition itself would be senseless. The point here is a new return into the domain itself, a return enriched with new knowledge and so with a new determination of the initial point and with a new delineation of an aim. The cyclic character of method is thus a tool of scientific progress.

This introduction enables us to give a more accurate characterization of strong, semistrong, weak, and semiweak methods.

A *strong method* Me_s is a method in which a set Z and all operations as well as sequences, eventually sequences on sequences, are precisely determined; they are single-valued and thus the whole domain D , and consequently the aim, too, are determined. Therefore

$$Me_s = \langle D, Z, {}^1\sigma_k, {}^1Po_l, {}^1PPo_1, V \rangle \tag{5}$$

It means that in (5), $PPo = f(Po_i)$, $i = 1, 2, \dots, l$, $k = 1, 2, \dots, n$; $l = 1, 2, \dots, m, \dots, < \aleph_0$.

We can say that in a strong method, the structure of the initial point determines the structure of the way. All strong methods have a corresponding algorithm which represents the method.

In a strong method there is a single-valued correspondence among V , Z and Po and therefore $V = f(Z, Po)$ or $V = Po_l(Z)$; eventually $V = PPo_1(Z)$.

Because in a strong method everything is potentially given, predictable, or calculable (deducible) already at the beginning, we cannot incorporate any *new* element without violating the whole process. It follows, among other things, that this method comprises the whole content and the whole truth given already at the beginning and therefore *there is*, in addition to other things, *no* place for incorporating either *hypotheses* or any other factors.

If the whole scientific method were to be characterized by a strong method, it could be probably expressed as follows:

An algorithm₁ → procedure according to an algorithm₁
 → limits of an algorithm₁ → an algorithmized introduction of
 (a new) algorithm₂ → procedure according to
 algorithm₂ → →

$$Me_w = \text{alg}_n(\dots (\text{alg}_1(Z)) \dots), \tag{6}$$

so that a scientific method Me_w would be an algorithm of algorithms.

In contemporary science there appear constantly more and more vigorous tendencies to change as many methods as possible into strong methods which would work in computers and thus replace to an increasing extent man's intellectual activity. These tendencies are made possible by the development of natural deduction in mathematical logic and by the development of computers.

A *semistrong method* Me_{ss} is a method in which an initial point, all operations on elements and all sequences are precisely determined. This

also determines D ; but the sequences are not single-valued, but not arbitrary multi-valued sequences either. They are limited by the possibilities determined by the character of the single-valued operations and by the precisely determined initial point.

$$Me_{ss} = \langle D, Z; {}^1\sigma_k, {}^jPo_l, V \rangle, \quad k = 1, 2, \dots, n, l = 1, 2, \dots, m \quad (7)$$

The best-known example of a semistrong method is the axiomatic method using the rule of detachment, i.e. the method which does not employ natural deduction. We know that what can be proved by means of the modus ponens in one way, can be proved in n -ways. By a semi-strong method we can reach an aim in various ways.

A semistrong method enables us to define all the elements of a set D (e.g. all theorems), but it does not unambiguously determine the way of a systematic natural acquisition of these elements. Therefore, the deduction of a certain statement which is known to be generally valid, is often difficult and can be found only after a long, random search. This means that a semistrong method is not linear, which does not prevent the *content* aspect of everything from being given at the very beginning; thus there is no place for the incorporation of a hypothesis with new content or of any new facts.

The sentence (7) defines only one type of a semistrong method. Other such methods can contain multi-operations but no multi-sequences. From this point of view we can speak of various degrees of strength (and of weakness, too) of semistrong methods.

A semistrong method that has multi-operations but no multi-sequences is stronger than a method having multi-operations as well as multi-sequences.

Strong and semistrong methods are methodologically *closed*. We know a great many closures; they can be realized by means of logical conclusions, mathematical calculations, various systemic relations, empirical implications, and the like. A methodological closure will be conceived with respect to operations and sequences. A method is closed in respect to an operation σ , if the following holds:

$$(\wedge a)(\vee b)(a \in D \wedge b = \sigma(a) \rightarrow b \in D) .$$

A similar condition can be formulated for sequences:

$$\wedge Po_i, \vee \sigma_j, Po_i \subset D, \quad \sigma_j = Po_i(\sigma_{j-1}) \rightarrow \sigma_j \subset D .$$

A methodological closure results in a closure of science. In other words, if science could be constructed by strong and semistrong methods, it would turn into a closed system itself.

In the history of science a conception of science as a closed system appears already in ancient Greece in the effort to create axiomatic systems which were considered as an ideal of science. In modern history such a conception is found in the work of Descartes and Leibniz.

Strong and semistrong methods are not the primary methods; for if the truth (and implicitly the whole content) is given at the beginning, then it must have been revealed by means of other methods, i.e. by means of weak and semiweak methods. Therefore we must consider strong and semistrong methods not only as instruments for the development of semiweak methods, but also as methods which use the results of semiweak methods.

A *weak method* does not have a systemic character and therefore either the operations or the sequences in it have no general character of closure; in the best case, the closure is only fragmentary or partial. Weak methods have neither a precisely defined domain D nor a precisely determined initial point Z . A sequence of steps is not always uninterrupted; some steps can be successively correlated, but they end many times in a blind alley. The research process has to return to the beginning and to start looking again for new directions eventually by other sequences. It means that in a weak method the relations among Z , Po and V are not single-valued, nor limitedly multi-valued, but they are independent from the beginning.

A weak method displays degrees of weakness too. An entirely descriptive method, which only registers and does not explain anything and therefore neither predicts nor foretells anything, is the weakest one. This method is so weak that there is no place for a hypothesis, i.e. for a preliminary design of an examined domain which is simultaneously a generalizing principle of prediction. Using this method, we become absolutely absorbed in the part of a domain which is just being examined, and we have no possibility of transcending it. If we realize that a hypothesis stands for a filter as well as for a principle of selection, i.e. a principle of demarcation of the essential from the inessential, then a weak method contains no place for a substance or a law, and therefore it does not enable one to make any distinction of the important from the unimportant, and so by means of it we can move only on the surface of reality. This method allows a constant influx of new knowledge in our arsenal — without enabling us to evaluate it.

A method of registering phenomena differs from a non-method only by the fact that the former originates in a conscious effort to investigate something "without any bias" (e.g. a domain), even though the means of the given method do not even enable us to create a complete picture of the domain under consideration. This means that a weak method does not unify systemically but only externally, e.g. by an existence of the domain itself.

A weak method enables only a construction of an absolutely empirical "science", a kind of the Plinius-like *historia naturalis*. This method forms a real beginning of a research into new fields and in that sense it must not be underestimated.

A method by means of which the phenomena are not only accumulated and registered but which also make it possible to search for relations valid for a given subdomain D_1 of a domain D , is a *less weak* method. Another step in a reinforcement of a weak method is a search for single-valued relations in the individual subdomains D_1, D_2, \dots, D_n of a domain D , and a search of single-valued relations between the investigated subdomains.

However, this procedure is also, to a great extent, a registrative one and a search for functions is *ex post*, after the objects are known in the individual domains.

The nearer we proceed to a unification of subdomains, the wider and the more single-valued are our possibilities of creating a complete view of this domain. This view also comprises possibilities of interpolation and extrapolation. We begin to treat a domain as a system. A weak method changes into a semiweak method.

A *semiweak method* $Me_{s,w}$ seems to be a mixture of a semistrong method and a weak method. A semiweak method is obtained if some basic features, e.g. the fact that everything must be determined from the outset, are removed from a (semi-)strong method or if some tightening conditions are added to a weak method. A strong method closes the sets D and M immediately at the beginning, whereas a weak method neither closes nor determines the set D (and even less the set M); in a semiweak method, as a kind of a synthesis of the above-mentioned methods, the sets D and M are closed and defined only *gradually*, i.e. the sets D and M are not given in it (even implicitly) but *are being formed*. This method neither anticipates nor leads to a definitely closed set but only to a kind of "incomplete" *open* set D . The set D becomes relatively closed at a certain stage of its development; it becomes a system (for example, an axiomatizable system), and is closed from the point of view of the

operations and sequences it develops in a systematic way (for example by deduction, calculations and algorithmisation, etc.). The success of this internal way makes it possible to apply the method to new domains and to extend its universe of applicability M . After reaching the possibilities and limitations of such a closed system, or by the occurrence of *new phenomena*, it opens up once more.

A semiweak method can be considered as a river which is channeled by regulation and "dams", and into which new knowledge constantly flows like new tributary streams; these dams get filled, broken and the river keeps on rolling in new river-beds; we regulate it again, erect new dams and the situation repeats itself—the river, the method pulsates. A semiweak method as a pulsating system consists of two parts: of a closing system (i.e. a stabilizing dam) and of an opening system. Thus not only D but also M is changing. These sets do not actually exist in a complete form. The semiweak method is in its substance temporal and historical. Therefore, a semiweak method can be characterized as follows:

$$Me_{sw} = \langle M_1, \dots, M_n; D_1, \dots, D_m; Z_1, \dots, Z_k; \sigma_1, \dots, \sigma_p; Po_1, \dots, Po_r; V_1, \dots, V_s \rangle; \quad (8)$$

therefore the indices $1, \dots, m; \dots; 1, \dots, s$ are temporal indices. This means that D_1, \dots, D_m are not subdomains of a fictitious domain D , but they temporally follow each other, i.e. D_1 is a domain of D known at a time t_1 , D_m is the same domain but deeper and more universally known at a time $t_m \cdot t_1 < t_m$. Z_1 , thus the initial point in a time t_1 contains a preliminary conception of a domain or of its parts which is presented in the form of a hypothesis. The operations σ and the sequences Po will be the tools for the development and the increased precision of the hypothesis, i.e. the tools for a more complete cognition of the examined domain. These are the closing operations leading to an aim V_i where V_i is simultaneously always a part of an initial point Z_{i+1} . This means that a semiweak method returns always from a previous aim V_i to the initial point and is therefore a cyclic method. In a semiweak method each aim is the only one, but it is preliminary, because from a relative character of our cognition it follows that our system of science must become open, otherwise it turns out to be insufficient for our comprehension of the new aspects of reality. These new aspects will be called N . Then the following holds:

$$D_i = Me_{sw}(D_{i-1}) \wedge N \neq Me_{sw}(D_i), \quad (9)$$

in other words, N opens D_1 and we continue the investigation so that

$$D_{i+1} = Me_{sw}(D_i, N) \quad (10)$$

$$V_{i+1} = Me_{sw}(V_i, N) \quad (11)$$

In this case the opening is such that ${}^N D_{i+1}$ implies that D_i^N is valid, but the contrary is false. This means that an operation of opening is not of a systemic character from the point of view of D_i , and so it is neither algorithmizable nor unambiguously definable. The operation of opening and especially the place and the way of opening cannot be incorporated into the initial conditions of the determination of D ; thus D cannot be closed in this way. This is the case because the limits of D_i and above all its extent are not known at the beginning. They are only partially known. Moreover, every finite sequence of a semiweak method as pertaining to the domains D_1, \dots, D_m can be or has to be changed *ex post* into a (semi-)strong method, i.e. every passed final part of a semiweak system can be rebuilt by means of a (semi-)strong method. If this were not the case, human cognition would change into a discontinuous mosaic without a unifying standpoint. The operations of opening are not completely unexpected and undefinable. There are several causes and possibilities of opening known from the history of science. They will not be examined here.

In conclusion, the question may be raised whether a scientific method is semiweak, i.e. whether science as a whole is or is not algorithmizable, or whether science can be constructed only by strong methods.

A scientific method as a whole cannot be an algorithm, for each algorithm is finite in such a way that there is a set of types of problems that are not solvable by a given algorithm though they can be solvable by another algorithm. Therefore if science and the scientific method were an algorithm, it would have to be a set of algorithms. The elements of this set (i.e. the algorithms) can be or cannot be connected algorithmically. In the latter case a scientific method as a whole would not be an algorithm. In the former case it would be an algorithm of algorithms. But also for this algorithm there would be problems unsolvable by it, and therefore, in order to compass science as well as the scientific method algorithmically, as an algorithm of algorithms, some absolutely universal algorithm should be taken into account. If only a single algorithm were omitted from this algorithm of algorithms, just this one might be needed for the solution of some scientific problem one day. At first glance, it could seem that a

universal algorithm would enable us to solve all the problems and so to define science and its methods appropriately. However, there exists no universal algorithm, because it would be internally logically contradictory: as an algorithm of all algorithms it should solve all problems but as an algorithm it cannot solve all the problems. This follows from the fact that the set of all recursive sets and functions is not recursive and that the set of all effective methods is not effective.

What is essential for this kind of algorithm is its closing function. Therefore it can never be adequately used for building science; in that case the future of science would be limited by its past and science could never break this limit. Such a science would in fact not have its own history — ahistoricism would prevail.

Leibniz's universal science, the *mathesis universalis* in its various historical forms remains only a pleasant dream or a directive: in science all that can, should be algorithmized.

This Page Intentionally Left Blank

IMPACT OF GLOBAL MODELLING ON MODERN METHODOLOGY OF SCIENCE

J.M. GVISHIANI

Institute for Systems Studies, Academy of Sciences of the USSR, Moscow, USSR

The topic of my paper suggests an answer to the questions as to what is global modelling, what are its specific logical and methodological features, and what contribution it can make and actually makes to the development of methodological knowledge. I would like to begin with a short historical digression.

Systems analysis and other simulation techniques have long gained wide acceptance in complicated problem solving. It was only in the early 1970s, however, that they found application in the studies of large-scale problems and processes evolving and running at a global level.

This was preceded by the creative efforts of a number of researchers and public figures concerned about the dissociation of scientific research and the global changes occurring in the socioeconomic and international life of the entire humanity. Many of them felt an imperative need for joint international research aimed at a thorough analysis of the global interrelationships in the contemporary world. It was necessary to think over the possibilities of setting up an international center for studying a complex range of world development problems.

It the late 1960s to early 1970s I happened to participate directly in the implementation of those ideas. Meetings with A. Peccei—a recognized expert in industrial management, M. Bundy—a former adviser to the President of the USA, F. Handler—President of the US National Academy of Sciences, and other prominent personalities made it possible to overcome some organizational barriers and reach consensus on the strategic questions of a new center for international research. The joint efforts resulted in the foundation of the International Institute for Applied Systems Analysis in 1972. It established scientific relations with the research institutions and centers of many countries of the world.

Somewhat earlier, in 1968, yet another international non-governmental

organization was set up. It was the Club of Rome, then headed by A. PECCEI (1984). Scholars, public figures, and representatives of business communities on the membership list of the Club concentrated on organizing research projects aimed at the study of the conflicting dynamics of world development. One of the major purposes of the research projects was to carry out a systems analysis of the difficulties mankind encounters at the current stage in its evolution. The founders of the Club of Rome proceeded from the premise that at present, when man is increasingly transforming the social and natural environment, it is no longer possible to set hopes blindly on the mechanisms of the global system's self-adjustment, but rather it is necessary to develop principles of world planning on the basis of general systems theory to control the complex dynamics of human activities within the context of his habitat.

A successful implementation of those plans was associated with construction and systems utilization of global models permitting the identification and analysis of major trends in the world development.

The first to suggest a concrete research project was H. Ozbekhan, a cybernetist who had attempted to respond constructively to the evergrowing complexity and uncertainty in the humanity's development. It was assumed that models representing the world dynamics and conducive to the identification of the major components of the system and relationships thereof would be constructed. It is worth noting that the task was not restricted to the systems analysis of the natural environment but involved a search for the normative values directly relating to social and political decisions. Unfortunately, the leaders of the Club of Rome failed to support the project due to its complexity and financial uncertainty: they were not sure of its practical feasibility in the near future.

Considering the time factor, the Club contacted J. Forrester, a prominent expert in mathematical modelling, who developed models using a systems dynamics method. In 1970 Forrester suggested a mathematical model "World-1" to simulate the global development and account for the interrelations of five variables—capital investment, population, food, availability of natural resources, and environmental pollution (FORRESTER 1971).

Following certain modifications and the construction of a model named "World-2", the leadership of the Club of Rome arrived at a decision on the relevant research. As was recommended by J. Forrester, the scientific and administrative responsibility for the project was vested in D. Meadows. In 1972 a group of young researchers headed by D. Meadows completed the mathematical computations on the model "World-3", and

published the research findings in the book *Limits to Growth* which set off stormy debates in the scientific communities all over the world (MEADOWS *et al.* 1972).

It is safe to say that this was actually the time a new branch of scientific research evolved, which is now referred to as global modelling. As a result of the ensuing discussion, the methodological principles of the research into the perspectives of mankind's development were reappraised, and a score of new models were developed, with every new model capitalising on the ever-increasing experience in global modelling.

Figure 1 presents the Global Modelling Universe. This diagram was first published within the framework of Major Programme I of UNESCO. We are reproducing it with some additions.

In the Mesarovic–Pestel model the world was regarded as a system of interrelated regions each with its specific features of change. In world models the object of analysis was the entire world. The Mesarovic–Pestel model was more flexible, opening the way for tackling the ill-formalizable problems. The Bariloche model (first discussed in 1974, described in press in 1976) was a reaction of Latin-American scientists headed by A.D. Herrera upon the questions that arose in the discussion of “World-3”. Bariloche’s modelling methodology is not designed to describe what the world will be, but what it could be, should all the decisions on distribution be aimed at a longer lifetime. In 1976 a group of Dutch scientists submitted for the consideration of the Club of Rome the MOIRA model dealing with the problem of nutrition. This model is based on neoclassic conceptions of economic development. In the model’s methodology econometric principles are widely used. In 1976 a group of scientists under P. Roberts, Systems Analysis Research Unit (SARU) of the Department of the Environment, UK, submitted for consideration a SARU model (SARUM) describing the long-term trends in the world development. SARUM was also designed to solve some methodological problems connected with understanding and overcoming the difficulties of the global modelling. The FUGI model developed by Japanese scientists was the first of the global models discussed in IIASA (1977). Such global models were based on the “input-output” method and macroeconomic dynamic modelling. In 1977 the United Nations World Model constructed by the research group headed by W. Leontief was also submitted for the consideration of IIASA. In the opinion of its authors this model was designed to investigate the impact of different problems and of the policies for their solution on the general strategy of world development (BRUCKMAN 1982, MEADOWS, RICHARDSON and BRUCKMAN 1982).

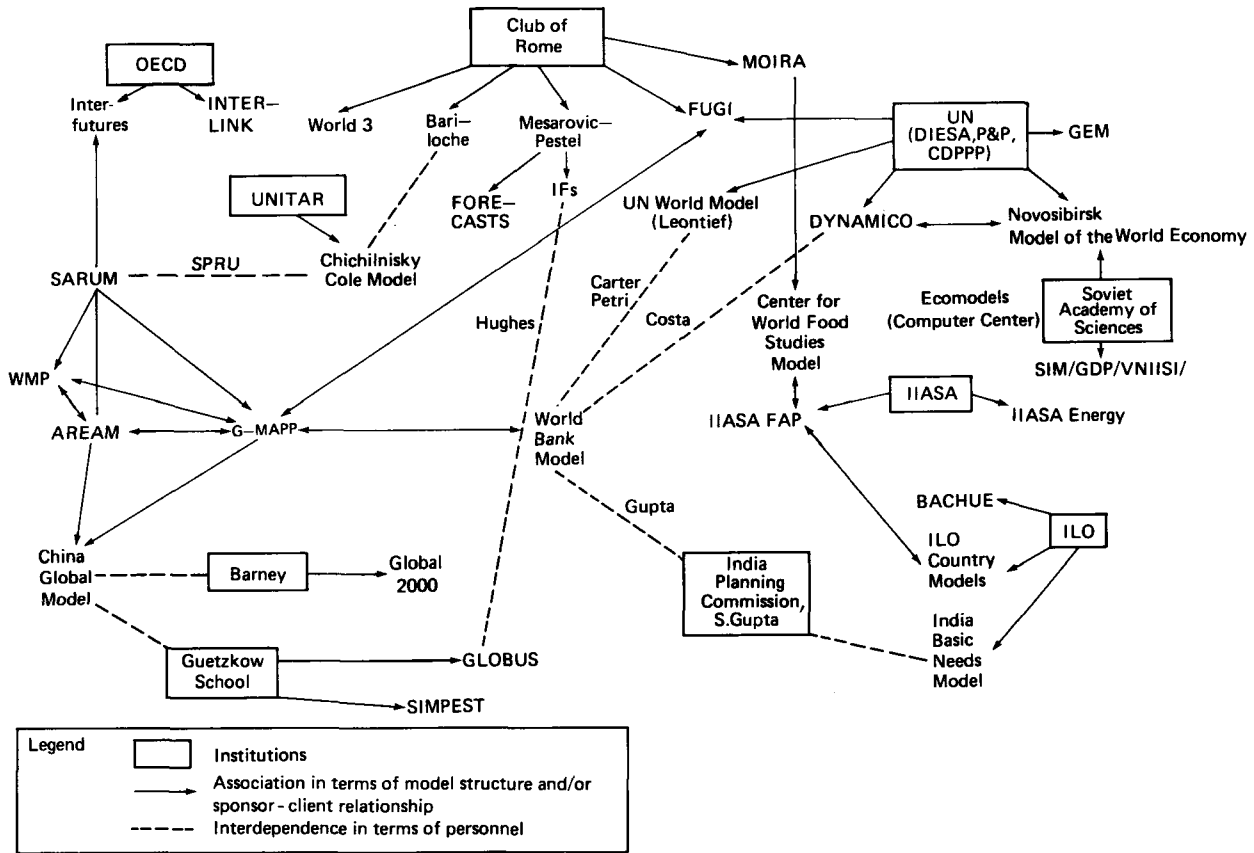


Fig. 1. The global modelling universe.

Research conducted in the Soviet Union by the Institute for Systems Studies, Moscow, within the limits of "Systems modelling and research of global and regional development processes" resulted in a further development of the conceptions and the methodology of the analysis of global development processes (*Elements of Man-Computer System for Modelling Global Development Processes* (VNIISI, Moscow) 1983, *Global Modelling: Social Processes* (VNIISI, Moscow) 1984, *Global Problems: Objective State and Assessment* (VNIISI, Moscow) 1986, *Sociological Aspects of Global Modelling* (VNIISI, Moscow) 1979, *Systems Modelling of Global Development Processes* (VNIISI, Moscow) 1980).

Global problems are universal and vitally important for all nations and peoples. It is only natural that separate uncoordinated efforts of even the most advanced states are not sufficient for their solution. That is why the problems are referred to as global.

The global problematic challenged contemporary science, an entire range of natural and social disciplines, including the logic and methodology of science. Due to its scale, overlapping of individual problems, influence of both objective and subjective factors, the global problematic call for a comprehensive approach to its analysis, considerably promoting the synthesis of sciences. At the same time the central part in the exploration of modern global problems is to be played by the social sciences. It is precisely these sciences that are called to provide us with a theory that is not only able to explain the causes and essence of conflicts in the social (in particular, technological) advance but also to outline the basic directions of the conscious and purposeful activities of peoples and states which would actually correspond to both the lofty humanitarian ideals and the entire process of the historical evolution of civilization (FROLOV 1987).

The global problems include at present those spheres of social development which feature the greatest aggravation of the conflict brought about by the current and projected situations, where the imbalanced growth and dysfunctional states cause or may generate catastrophic implications in the foreseeable future. The first findings of global research were so alarming that they bordered on apocalyptic prophecies. It would be wrong, however, to speak about the exaggerated gravity of global problems, for one cannot remain calm when more than one billion people on the Earth are deprived of the bare essentials, are on the brink of survival, and the data for the last decade indicate that the availability of food in many developing countries has considerably reduced.

An important role is played here by methodological problems. Indeed,

the choice of an adequate methodology has a strong impact on the solution of the problems related to the management of vital issues. This is an immense task of actually worldwide historical significance. In fact, earlier separate countries did make attempts, with varying degrees of success, to control the man-nature interaction processes, for example; but at the level of the entire global system those processes ran spontaneously. A cardinal methodological question arises here however: is it possible in principle to introduce reasonable controls, adjustable influences in the global system dynamics? We believe there is such an objective possibility. Should we abstract ourselves from political factors, we will realize that the essence of the methodological aspects of the global problematique is associated with systems methodology. This is related primarily to the fundamental fact that the very global problems form a complex dynamic system. Let us consider in this connection a list of global problems (FROLOV 1987, ZAGLADIN and FROLOV 1981). The most acknowledged of them are:

- new world war prevention, curbing the arms race, comprehensive and full disarmament, creation of a non-nuclear, non-violent world;
- final elimination of colonialism, neo-colonialism, racism, apartheid, all forms of oppression, discrimination and inequality;
- implementation of the right of all nations for free and independent development, overcoming backwardness of the developing countries;
- reconstruction of international economic relations on a just, democratic basis, establishing a new economic order, ensuring world economic security;
- use of the achievements of scientific progress for the benefit of mankind;
- all-round comprehensive development of man as the highest aim of the social progress, guaranteed implementation of the human rights;
- democratization of education and elimination of illiteracy;
- ensuring demographic dynamics balanced with the development of the forces of production;
- ensuring food supply needed for the rapidly increasing population of the planet;
- environment protection, curbing the growth of pollution caused by man's activity;
- rational use and increase of the mankind's energy and raw materials potential, elimination of the shortages of traditional energy resources and fresh water, raw materials depletion;

- World Ocean resources exploitation;
- struggle with dangerous diseases, provision of the medical care for the entire world population;
- elimination of the imbalance in international information exchange, creation of new international information order;
- preservation of world cultural legacy protection, provision of access to the cultural values for everyone.

The suggested ordering of global problems is certainly arbitrary. What is really essential is the very mention made of the areas where the aggravation of the situation is quite evident. This traditional classification of global problems bred illusions in some specialists about the possibility of an isolated analysis and solution of each problem. Thus, there appeared numerous studies of the impact of the processes related to some global problem, and of the characteristics of the processes inherent in one or several global situations. Such disconnected chains of causal relations usually identify the forthcoming implications rather than distant perspectives. They lead to rather trivial conclusions and often to fruitless recommendations. Suffice it to mention an idea shared by some international organizations, according to which the major misfortunes of the developing countries are due to an excessive population growth, which allegedly brings to naught all the efforts for the increased food production, consumes all available resources, and prevents these countries from getting to a higher level of development. Hence, an inference is made from here that the developing countries must pursue an active demographic policy. And since the mechanisms which form the attitude toward population reproduction and affect its variation, were ignored by the researchers, an active demographic policy was confined just to birth control.

Yet another example of a one-sided approach to the solution of global problems, alien to systems ideas, is the position of those researchers who can see a way out of the food problems suffered by the developing countries in an aggregate application of approaches. It implies an increased food aid to these countries and introduction of modern agricultural technologies, including utilization of bumper crops and fertilizers, herbicides, and insecticides. Such an approach has proved to be patently inoperational. As the "Green revolution" experience showed, this strategy, attractive as it may be, does not provide for the food problem solution. On the contrary, it breeds a host of unforeseeable negative implications of socio-economic nature.

Individual partial solutions to large-scale social problems, such as attempts to change the distribution mechanism without restructuring the productive relations, are doomed to failure.

Now let us consider the specific logical and methodological features of global modelling.

Global modelling is a branch of systems global modelling, which makes an important part of the modern methodology of science (GVISHIANI 1985). Indeed, in modern knowledge an increasing significance is acquired by systems modification of research methodology of science in general, where the traditional use of models in mechanics and physics is further developed to raise it to a higher level (BLAUGER, SADOVSKY and YUDIN 1977). This became possible due to the following recent breakthroughs in various fields of knowledge and practice:

in methodology — intensive development, beginning from the middle of the century, of systems research, in particular, cybernetics and mathematical methods like systems dynamics;

in technology — the development of computers capable of processing huge volumes of information and automating a lot of time-consuming routine procedures required in the modelling of systems with complex organization;

in social reality — the need to operate with exceedingly complex systems and problems characterizing the era of technological progress, in particular, the exceedingly complex, multi-attribute character of the socio-technical and socio-natural systems underlying modern civilization.

In order to apply systems analysis to such complex systems, it was necessary to move from the traditional “physical and technical” modelling to that based on systems methodology, where the emphasis is placed on a formalized representation of a certain whole consisting of heterogeneous parts. Such representation is made possible by the capability of modern computers, complemented by non-formalized (descriptive) expert assessments (LAPIN 1984).

Contemporary systems modelling in a broad sense of the term is a totality of simulation techniques based in a varying degree on systems methodology and computer technology. The most popular approach to this area today is a computing experiment in the form of man-machine interaction (GELOVANI 1985).

Four dimensions of the general methodology of knowledge underlie the basic cognitive function of global modelling:

1. Holistic representation of the simulated system.

2. Reasonable approximation of the original system by the model representation.
3. Projection into the system's future (extrapolation).
4. Interdisciplinary synthesis, a comprehensive analysis of a real complex system or of a complicated problem by different interacting sciences working toward a single goal.

The system of methodological principles of global modelling is structured on three levels: (1) general philosophical level; (2) methodological level; (3) science-specific level.

The general philosophical principles of global modelling are:

- the laws and categories of dialectics;
- the principle of the unity of society and nature;
- the principle of wholeness;
- the principle of humanistic development.

We believe that the general philosophical principles, primarily the principle of development and the principle of dialectics of the opposites, etc. serve as a key to global modelling. In their turn they embrace the following principles:

— *Principle of the unity of society and nature.* In the cultural development of the civilization we may observe a change of the paradigm from the attitude of conquering the nature, considered as a certain passive object of human action, to the attitude of achieving a reasonable compromise between society with its technosphere and the objective capability of the biosphere.

— *Principle of wholeness.* This principle refers to all elements of global modelling: the object of modelling, the data bank of the world processes, the criteria by which the alternatives evolving in the modelling are compared.

— *Principle of humanistic development.* The global system cannot be treated either as unchanging or as inevitably degenerating. Its evolution proceeds in such a way as to orient its progress towards man; his comprehensive perfection.

The methodological principles of global modelling are:

- the systems principle;
- the principle of interdisciplinarity;

- the principle of the unity of the subjective and the objective;
- the multimodel principle;
- the principle of interaction;
- the principle of dynamism.

These principles can be characterized as follows:

Systems principle: The object of study is to be represented as an integrated whole whose dynamic behaviour is a result of the interaction of its parts, in the endeavour to adapt to the changing environment.

Principle of interdisciplinarity: Complex systems are, as a rule, connected with the subject's actions and processes stemming from different forms of the movement of matter. Therefore even the definition of a problem in global modelling presupposes the integration of knowledge.

The principle of the unity of the subjective and the objective may be characterized in the same way as the principle of the unity of society and nature above.

According to the multimodel principle, systems modelling requires a variety of models. Each model can provide its particular image of the object. The interrelationships between these various models of a single object within the framework of the modelling system may be quite different — correspondence, complementarity and even confrontation.

Principle of interaction: Systems modelling not only represents reality but, what is more important, provides a basis for the formulation of a set of alternatives. The selection of the optimal alternative is made by the researcher who interacts with the computer. The possibility of an interactive process accounts for the dynamism of systems modelling methods of knowledge.

Principle of dynamism: Flexibility and dynamism have to be inherent in any type of modelling. In systems modelling they take a variety of new and effective forms, which determine to a large degree the scale of research capability and the range of its applications.

A qualitatively new stage is heralded by the advancement of computer software and algorithms, in particular, the higher-level languages facilitating the researcher's interaction with a computer model. This is conducive to a rapid restructuring of the model, replacement of separate blocks, and a modification of initial conditions and variables.

The combination of dynamism and multimodel principle makes it possible to employ a structured set of models. The latter can partly compensate for a certain limitation of computer experiment as compared to the traditional analytical methods of mathematical modelling that do

not require numerical values of all the model variables, and allow the subject to take a wider look at the complex object under investigation or a complex multi-factor problem.

Understood in this way, systems modelling may also be synonymous with modern modelling research — from the “entropy” models of traffic flows and mathematical models in economics to the semi-qualitative methods used in ill-structured problems of choice. However, the specific systemic elements of these methods may differ and may refer to different aspects of research.

If we pass to a more detailed level of special modelling methodology, we may identify the following specific principles of systemicity in global models:

- interdisciplinary preliminary systems analysis of the object under study, with the aim of delineating the relevant parameters and characteristics as well as the essential relations between them;

- identification of those parameters and characteristics that cannot be the subject of formalization; application of expert systems techniques to the analysis of this class of problems;

- application of formal methods, in particular, those of mathematical modelling, the creation of a formalized mathematical model of the given structure of elements and relations;

- computing experiment: an experiment on a mathematical model for different values of the parameters it contains with the help of computing methods;

- interactivity, the dialogue-like character of modelling which allows to implement the unity of the formal and non-formal;

- interaction of modelling, allowing model correction, perfection and extension.

Of great significance is also the ontological basis of the specific epistemological features of global modelling stemming from the character of the objects and tasks of systems modelling. Among them the complexity of the investigated processes and phenomena seems to be the most general notion characterizing the subject matter of systems research. At an early stage of the development of cybernetics, von Neumann and Kolmogorov recognized the significance of this measure of reality. It accounts for several epistemological features of systems modelling, including a new qualitative stage of research. The model has always been a peculiar factor, a sort of “quasi-object” instrumental in the cognition of the object. In systems modelling we observe a transition to a “second-

level modelling” as the models themselves are so complex that “models of models” must be built (i.e. research inevitably becomes a multi-stage process). The same tendency is observed when more often than not a natural experiment is replaced by a machine experiment.

It is common knowledge that complexity may express itself in a variety of models. Accordingly, systems modelling has developed several alternative approaches to coping with it. I want to stress once more that one of these ways is the tendency to represent the object under investigation as a set of models that may be complementary to one another. Another way—approximation—has always been an indispensable element of natural science but was always regarded as a “forced”, “tactical” means. The study of complex systems in various fields of science has radically changed the attitude to approximation: it has been “legalized”, but it has not got the status of a methodological norm of research.

Many systems models have an important point of difference in comparison with the traditional monodisciplinary models that result in fact from an application of a scientific theory to the range of phenomena covered. This point of difference is problem-orientation. Earlier the importance of problem- (not object-) orientation was pointed out by V.I. Vernadsky. Even though the identification and definition of problems cannot but depend on the aggregate knowledge, their content is mainly determined in practical activity. Therefore the model has to take into account various factors pertaining to different scientific disciplines, i.e. it must perform (on the level of a model) an interdisciplinary synthesis. It is noteworthy that a methodological pre-requisite of such synthesis is a universal form of knowledge possessing integrative properties (by the degree of commonality—dialectics, mathematics, systems cybernetic methodology). The interdisciplinary modelling synthesis is most fully employed in the systems modelling of global and regional development.

In addition to the complexity of the global subsystems, the global model deals with the complexity of goals, motivations, and attitudes underlying human activity. While the central goal of classical science is to search for universal laws, systems modelling strives, in addition, to learn and control the goal-oriented and unique processes occurring in the complex objects of investigation. The systems model is a characteristic cognitive instrument of systems research integrating the goal-oriented and cause-oriented description of socio-technical systems, where there is no clear demarcation line between the object and the subject, and which greatly depend on the objective laws of existence as well as the strategy and tactics of purpose-oriented human behaviour. By means of modelling

one may bring together the natural universal features and the specific features of this or that technical or social project. Therefore models increasingly become models of action (not models of things), instruments of the assessment and choice of control leverage.

Action and control in the modelled systems cannot be adequately reproduced by means of formalized models alone, since man is their organic part, their subject. Therefore the integration of the formalized and non-formalized modelling components into an integrated system is an important tendency of systems modelling. An indispensable part of this system is the researcher (an expert, a decision-maker) who is able to interpret the results of the formalized modelling in a meaningful way. The conceptual interpretation which supplements the description of processes and problems in "objective" categories by an account of their meaning, becomes indispensable in systems research that includes the elements of the society. The most effective modes to organize such systems are the interactive man-machine complexes.

The interaction between computer models reflects most clearly the link between systems modelling and the subject/object relationship dynamics.

In the concluding part of my paper I would like to consider the complex problems which global modelling encounters in the analysis of social and cultural processes. Some of them relate to the still uncertain, evolving status of global modelling in the structure of modern science, others to the internal problems of modern social sciences.

Global modelling has dashed into the domain of science with bold claims to the analysis and even solution of the problems that traditional science did not have the courage to handle. The first fearless steps were made by small groups of scientists, where one often could not find any familiar, prominent name. In their stead there were semi-professionals, management practitioners, and scientists engaged in applied science whose judgments of the problems were somewhat simplified and naive. This gave birth to alarm and scepticism.

However, global modelling has already made its contribution to modern science: it proposed, constructed and imposed on science its specific object of investigation—a global system that is not part of any larger system, and does not have other systems of the same order of magnitude as objects of comparison.

This essential feature of the object makes its analysis difficult. Besides, it calls for non-traditional approaches, and makes it necessary to search for new methodologies. In the process global modelling may acquire a new role in cognitive activity. For example, lack of any higher systems or

systems of the same order means that the goals of the global system can be represented in the model only in value categories. The traditional attempt in science to place the values outside the research boundaries turns out to be impossible with respect to global modelling. The values of the researcher who is building the model are found to be its indispensable element. The model becomes a specific mode of reflexion of culture, rational comprehension and re-comprehension of cultural values within the nature/science paradigm (LEIBIN 1982).

This process can be observed in the existing western global models. The language of their values is the language of a certain culture, and its notions are economic growth, intellectual rule of science and technology, human needs, etc. However, in a scientific context these stereotypes of ordinary consciousness become the scientifically postulated goal of the global system's development, and acquire a new meaning. It turns out that they need to be substantiated and that is usually a symptom of the re-evaluation of values.

Performing this useful reflexive function, global modelling finds itself in the focus of one of the most subtle problems of science — its attitude to values. Here, owing to the specific features of its research task and method, global modelling can make its contribution. In principle it cannot free itself from values, but neither can it accept the values as a “full-fledged” element of scientific knowledge. It is this particular statement that such prominent representatives of western social science as A. Gouldner and J. Galtung stand for (GALTUNG 1980).

They do not consider the new social science, say “reflexive sociology”, to be objective knowledge that is entitled to any assumptions and conclusions beyond science insofar as they are identified and understood. Gouldner compares social science with a biased man. He believes that a social science researcher deals with two equal types of reality — with the facts obtained by research and “personal reality” (what the researcher has seen, heard or experienced).

However, despite its load of values, global modelling cannot take this stand. Its major goals — forecasting, analysis of the alternatives for global development, and humanistically responsible warnings pre-suppose a convincing argumentation, the language of universality, i.e. objectivity, rationality, and logic, which is incompatible with appeals to subjective assessment and unique experience.

Therefore global modellers are obliged to treat values as knowledge and actively utilize them in different forms and on different levels.

The global modelling community (first of all, the alarmists) give a

special attention to the process of general values evolution. It clarifies and explains such phenomena as the non-rational and non-functional lagging of value consciousness, which is now manifest with respect to ecological values or relics of culture, or the close relationship of values and direct perception. This latter lies at the bottom of the (until recently) careless attitude to radiation that lacks colour, smell, and heat.

Global modelling made clear a need for the research tools for exercising scientific control on the value elements of analysis. Various methodologies and techniques of defining the problems to be investigated and solved by means of such models are already available.

The Institute for Systems Studies (VNIISI) is engaged in elaborating an approach to the formulation of social problems as a specific manner of organizing knowledge. It is based on the unity of analysis of the objective and subjective processes going on in human society. It is no less important to comprehend scientifically such a value-based process as the interpretation of the modelling results. The relevant VNIISI research in this direction is oriented towards designing a system of theoretical and methodological links between the model's outputs and the interpretative description, beginning with the block of formalized interpretation and ending with the assessment of alternatives from the viewpoint of generalized socio-historical criteria. An especially complex task is to make a typology of values on the basis of their role and functions in societal development and, consequently, in the investigation of this development (*Global Modelling: Social Processes* (VNIISI, Moscow) 1984, *Global Problems: Objective State and Assessment* (VNIISI, Moscow) 1986, *Sociological Aspects of Global Modelling* (VNIISI, Moscow) 1979).

The proper scientific problems arising from the analysis of social processes by means of global modelling are also sufficiently complex. How can one investigate an object if there is nothing that resembles it? So far the problems of research are solved here empirically or theoretically.

There are few, if any at all, dynamic rows of indicators for the social processes required in modelling. Therefore researchers are forced to employ the methodology of cross-sectional research, where the data on countries that are on different levels of economic development are treated as the data on a single country that has passed all these stages. The modern global system is identified with a generalized country existing only in theory. This country began its march from the per capita annual income of less than 100 dollars to a level of 3000 dollars, with the Gini scatter coefficient of 30–60 that is characteristic of social inequality. But

this approach is not correct as it does not take into account historical nor contemporary differences in social, cultural and economic structures or regional community.

Another difficulty lies in that the modelling of social processes in general, and in global projects in particular, is not yet independent from a purely methodological point of view. Social variables or indicators are so far determined as if they were the social consequences of economic processes. It means that the analysis of social processes cannot yet rely on the original (not "secondary") models (*Sociological Aspects of Global Modelling* (VNIISI, Moscow) 1979).

Nevertheless, a model of social processes is needed not only to trace the dynamics of "social indicators", but also to introduce into the system the mechanism of socio-historical movement, i.e. to model man's social activity as a conscious, purposeful and socially regulated social behaviour (*Global Modelling: Social Processes* (VNIISI, Moscow) 1984).

Man as the subject of social processes and social development does not possess any certainty that can be expressed in a finite number of end-parameters. Therefore the goal and, consequently, the mechanism of social development cannot be formulated by means of describing man's goal-state or the satisfaction of his needs. This is what was meant by K. Marx when he spoke of developing all human faculties as such, irrespective of any pre-fixed gauge.

Man is identical with man's world. Therefore to model man as the subject means modelling social, economic, cultural and other processes. Thus, the circle is closed when they model man's social activity as individual behavior and state (*Global Problems: Objective State and Assessment* (VNIISI, Moscow) 1986).

Consequently, there is a need for the analysis of more stable, independent fundamental social processes possessing independent and "modellable" dynamics, the objective logic of movement. No doubt, the most important of these processes is the social structurization of society. It performs serious functions relating to the preservation and development of the social system — therefore it cannot have "dependent" dynamics. The internal logic guiding the regulated and non-regulated social differentiation and the emergence of new social structures are thoroughly investigated within the framework of global system's modelling in VNIISI. The attention is focused on the study of such currently important properties of social structures as flexibility and restructuring capability, without involving any serious social or human costs (*Global Modelling: Social Processes* (VNIISI, Moscow) 1984).

There is a need for active search in another area too. To model social processes, one will need a universal gauge, an objective quantitative measure such as, e.g. money for economic processes. The approach employed by VNIISI suggests social time as such a universal measure. Theoretical research in this line is aimed at defining its specific features compared to physical, biological and subjective time. Empirical modelling currently covers only the distribution and dynamics of society's working and leisure time.

Global models forecasting attempts are our reflexion of the future. And the conception of the future requires an alloy of science and intuition, scientific and value analysis, moral assessments and artistic images. Its reliability depends on how organically these modes of the intellectual conception of reality can be integrated. Modern philosophy requires new levels of thinking which extend the temporal and cultural boundaries of its traditional problems, plus a more profound knowledge of global problems and the alternatives for their solutions.

References

- BLAUBERG, I.V., SADOVSKY, V.N. and YUDIN, E.Y., 1977, *Systems Theory: Philosophical and Methodological Problems* (Moscow).
- BRUCKMAN, G., 1982, *Les modèles mondiaux*, *Futuribles* 59, pp. 17–30.
- Elements of Man-computer System for Modelling Global Development Processes*, 1983, Coll. papers, vol. 3 (VNIISI, Moscow, in Russian).
- FORRESTER, J.W., 1971, *World Dynamics* (Wright-Allen Press, Inc., Cambridge, Mass.).
- FROLOV, I.T. (ed.), 1987, *Socialism and the Progress of Mankind. Global Problems of Civilization* (Moscow, in Russian).
- GALTUNG, J., 1980, *The True Worlds* (N.Y.).
- GELOVANI, V.A., 1985, *A man-machine simulation system for global development processes*, in: Gvishiani, J.M. (ed.), *Systems Research II, Methodological Problems* (Pergamon Press, Oxford).
- Global Modelling: Social Processes*, 1984, Seminar Proc. (VNIISI, Moscow, in Russian).
- Global Problems: Objective State and Assessment*, 1986, Seminar Proc. (VNIISI, Moscow, in Russian).
- GVISHIANI, J.M., 1985, *Theoretical and methodological foundations of systems research and the study of global development problems*, in: Gvishiani, J.M. (ed.), *Systems Research II, Methodological Problems* (Pergamon Press, Oxford).
- LAPIN, N.I., 1984, *Nonformalised components of modelling system*, in: Gvishiani, J.M. (ed.), *Systems Research I. Methodological Problems* (Pergamon Press, Oxford).
- LEIBIN, V.M., 1982, *Models of World and Vision of Man* (Moscow, in Russian).
- MEADOWS, D., et al., 1972, *The Limits to Growth* (Earth Island, London).
- MEADOWS, D., RICHARDSON, J. and BRUCKMAN, G., 1982, *Groping in the Dark: The First Decade of Global Modelling* (J. Wiley, Chichester).

- PECCEI, A., 1984, *Le Club de Rome: ordre des jours pour la fin du siècle*, *Futuribles* 76, pp. 3–14.
- Sociological Aspects of Global Modelling*, 1979, Coll. papers, vol. 6 (VNIISI, Moscow, in Russian).
- Systems Modelling of Global Development Processes*, 1980, Coll. papers, vol. 14 (VNIISI, Moscow, in Russian).
- ZAGLADIN, V.V., and FROLOV, I.T., 1981, *Global Problems of Today's World: Scientific and Social Aspects* (Moscow, in Russian).

CONCEPTUAL CHANGE AND THE PROGRESS OF SCIENCE¹

DAVID PEARCE

Institut für Philosophie, Freie Universität Berlin

Introduction

The first edition of Thomas Kuhn's book *The Structure of Scientific Revolutions (SSR)* was published 25 years ago, in 1962. Since then, it has exerted a profound and enduring influence on philosophers of science. It has also, more than any other work, transcended professional boundaries and helped to shape the informed image of science that we have today. At the same time, its impact has been highly controversial, provoking some of the worst disagreements and the most heated debates among philosophers as to the character of scientific knowledge and its patterns of growth.

Following Kuhn's attack on traditional empiricist and critical rationalist epistemologies of science, most philosophers of science have lived out an uneasy compromise between the need to take the history of science seriously and the desire to avoid the kind of historical relativism, and even irrationalism, that has been associated, rightly or wrongly, with Kuhn's philosophy. In coming to terms with this compromise, philosophical conceptions of "scientific progress" have been successively liberalised, relativised and generalised. Today, we rarely view scientific hypotheses, laws or theories in isolation, as items to be confronted directly with Nature or evidence. We prefer to analyse progress with reference to larger units—paradigms, research programmes, research traditions—replete with their associated background assumptions, aims and methodological norms. We also tend to look at progress less as an

¹ Aside from minor changes, this paper is the text of the lecture delivered at the LMPS Congress in Moscow, August 1987. I am grateful to Ilkka Niiniluoto for helpful comments.

absolute than as a relativised concept: what counts, at bottom, is whether a theory is *more progressive* than its rivals. And we are less concerned to make the progress of science seem cumulative and fully rational in every respect and at every stage of its development, being content to have science globally and in the long term progressive and rational.

In the light of Kuhn's work and its subsequent reception, I want to discuss some aspects of the problem of progress and conceptual change in science. In particular, I shall concentrate on what is perhaps the most renowned concept of Kuhn's book: "incommensurability". This term is nowadays a catchword for so many different notions that it has to be used with caution. However, there is only one sense that really need concern us, where progress and conceptual change are at issue: the sense in which two scientific theories or conceptual frameworks may be said to be incommensurable if, or because, there is no adequate translation from the language of one into the language of the other. This is the sense of incommensurability in which the matter of conceptual change becomes a problem in the philosophy of science and where it raises difficulties for our analysis of progress.

I shall treat four themes related to this problem. First, I shall argue that attempts to by-pass matters of conceptual disparity and incommensurability when describing scientific progress have been unsuccessful. Secondly, I shall discuss and reject some proposals to regard (in)commensurability as an exact, logically defined concept. Thirdly, I shall claim that Kuhn's recent arguments for the incommensurability of specific theories are still unpersuasive. And, fourthly, I shall suggest that the account of theory structure and development that Veikko Rantala and I have offered provides a viable way of integrating conceptual change into the analysis of scientific progress.²

1. Progress and conceptual change

Conceptual change in such a fundamental feature of the growth of scientific knowledge that it is sometimes hard to understand how it should ever have come to be a *problem* of special concern in the philosophy of science. It seems so natural to say that science makes progress *through* conceptual change, by refining, transforming and inventing concepts, by

² Many of the issues raised in this paper are dealt with at considerably greater length in PEARCE (1987).

creating new languages along with new theories. Yet one often has the impression that many philosophers of science would be more inclined to say that science makes progress *despite* conceptual change; as if the need to modify the *language* of science should in some sense be thought to be a weakness or a failure of our process of knowledge acquisition, instead of being a positive advantage; as if *conceptual* creativity, so highly praised in the arts and in other walks of life, should be regarded with suspicion in the sphere of the natural and the social sciences.

Puzzling as this view might at first sight appear, it actually has many adherents: philosophers who hold that even if a certain scientific concept acquires a new definition, new contexts of application, or new methods of measurement leading to different results, still the concept has not really changed its meaning. When, with reluctance, philosophers have conceded that scientists do, at least sometimes, change their concepts, it is often maintained that nevertheless they are really still talking about the same *things*. To prove just this, Hilary Putnam invented (along with Saul Kripke) one of the most remarkable philosophical “white elephants” of recent times: the causal theory of reference.³ Others, like Hartry Field, have taken the matter a stage further, arguing that even if scientists who hold different theories are perhaps not always talking about precisely the same things, they must at least be talking about *partly* the same things. What it means to refer only partly to something is of course in need of some clarification, and, to help out, Field proposed what he thinks to be a fundamentally new semantical theory, that of partial reference.⁴

Without wishing to imply that no interesting results have emerged from recent philosophical discussions of meaning and reference, I fully concur with Dudley Shapere’s remark that “the technical concepts of meaning and reference stemming from the philosophy of language have failed to clarify the scientific enterprise”.⁵ I also think we should heed his request to “exorcise completely the error of supposing that scientific reasoning is subservient to certain alleged necessities of language, and that the study of the latter is therefore deeper than the study of the former”.⁶ As far as the analysis of scientific progress is concerned, my own plea is that we should start to take conceptual change seriously. That is to say, conceptual change should be recognised for what it is: a basic ingredient of the

³ See especially PUTNAM (1973, 1977).

⁴ FIELD (1973).

⁵ SHAPER (1982).

⁶ *ibid.*

growth of science. And our models and rational reconstructions of that growth should assign a prominent place to the changes which the language of science undergoes.

The uneasy compromise I referred to, which has disquieted many philosophers since Kuhn, has often emerged in the following form. To respond to Kuhn's challenge and eliminate any suspicion of irrationalism or relativism about science, we must, they say, on the one hand, attend to the real historical development of science, on the other hand, show that meanings, or at least references, or at least partial references are preserved under changes of theory. The inference seems to be that the whole problem of incommensurability, in its turn a problem of non-translatability, arises because meanings or some other semantical features of language are alleged not to be historically stable. If one can show, by example or by appeal to philosophical theory, that some kind of semantic stability is defensible, then — the argument continues — Kuhn's thesis of meaning variance, or at least his thesis of incommensurability, would be undermined.

This may help to explain the curious tendency among philosophers to play down conceptual change and deny scientists their full rights to linguistic creativity. However, the line of argument just mentioned is patently defective: continuity of meaning or of reference is neither a necessary nor a sufficient condition for commensurability. It is not necessary because what is required of a translation is that it re-expresses in one language what is said or written in another. In other words, a translation maps sentences to sentences having the same or nearly the same semantic values. The meanings of individual words (even shared words, if they occur) in two languages may be completely different, but, as long as the target language is rich enough to express statements made in the source language, then translation may be possible.⁷ The condition is also not, by itself, sufficient. The reason is that to be commensurate, two theories must articulate *claims* which, possibly mediated by translation, stand in some sort of logical contact to one another; for instance, they might offer conflicting or compatible solutions to some cognitive problem, or rival explanations of some fact. Their mere sharing of some

⁷ I am assuming here that preservation of sentence-meaning is the principal adequacy criterion for translation. In cases where we have a recursively specified syntax and a principle of (meaning) compositionality (typical of most of the formal languages and semantics devised for mathematics and science), this criterion may be re-expressed in terms of smaller lexical units (atoms), along with the requirement that translation be recursive.

particular concept is thus not, by and of itself, an indication that two theories formulate commensurable claims involving that concept.⁸

A growing number of philosophers has joined Kuhn in acknowledging the presence (or conceding the possibility) of incommensurable theories in science, whilst insisting, against Kuhn, that the usual categories of rational theory appraisal and selection still apply. They include instrumentalists, like Larry Laudan and Wolfgang Stegmüller, as well as realist-materialists, like Paul Churchland, Geoffrey Hellman and Frank Thompson.⁹ For these writers, therefore, translatability is not a necessary prerequisite of rational choice. However, none of them, in my view, presents a convincing case for this thesis. Laudan, for instance, argues that two scientific research traditions could be rationally compared for their problem solving effectiveness (PSE) even if they are “utterly incommensurable in terms of the substantive claims they make about the world”.¹⁰ But Laudan overlooks the fact that, according to his own analysis, PSE is a highly *comparative* notion, relying crucially on the fact that rival theories or research traditions not only share problems but also assign them a roughly similar importance or *weight*. Thus, if two theories are, in Laudan’s sense, genuinely incommensurable in that they share no common problems,¹¹ a comparative assessment of their respective PSEs can hardly provide an index of rational choice between them.

The other writers I mentioned try to express the rational comparability of incommensurable theories in terms of traditional concepts like *reduction* (or, in the case of Hellman and Thompson, a slightly weaker relation of *determination*). In Churchland’s analysis, however, the concept of

⁸ Except perhaps in pathological cases, commensurability should result, however, when there is complete continuity of meaning in the sense that two theories share all their concepts.

⁹ HELLMAN and THOMPSON (1975, 1977) do not so directly concern themselves with problems of scientific *progress*; but they do deal, quite specifically, with questions of conceptual change, theoretical equivalence and incommensurability. I include them here because they formulate theses which, if correct, would have powerful implications for our analysis of the dynamics of science. One of their claims, in particular (see below), is remarkably similar in content and motivation (though obviously independent in its origins) to what I call Stegmüller’s *rationality thesis* (note 14 below). I discuss several aspects of Hellman and Thompson’s physicalist materialism in PEARCE (1985).

¹⁰ LAUDAN (1977), p. 146 (original in italics).

¹¹ The sharing of common problems is Laudan’s own yardstick for the commensurability of theories; see LAUDAN (1977: Chap. 4). Actually, in Laudan’s claim quoted above, “incommensurable” can be understood both in the usual sense of “non-translatable” as well as in Laudan’s special sense of “problem-disjointedness”. Both readings are analysed in PEARCE (1987: Chap. 3).

reduction becomes weakened almost beyond recognition, so that it is unclear from his account how a reduction between untranslatable theories is really to establish a cognitive connection of the required kind. At the same time, his discussion of examples like that of classical versus relativistic mechanics fails, in my view, to sustain the thesis of untranslatability.¹² The more sophisticated concept of reduction employed by Stegmüller and the structuralist school is (like the Hellman–Thompson concept of determination) defined purely structurally as a certain kind of mapping (or relation) between the models of two theories. It can be shown that, under plausible conditions usually fulfilled in cases of scientific reduction, this kind of structural or semantic mapping induces a syntactic or translational mapping between the languages of the theories in question. Hence, reduction in Stegmüller's sense implies translation, and even establishes a relation of (generalised) logical inference between the theories' laws.¹³ His thesis that one can have reduction without linguistic commensurability is thus, I would argue, untenable.¹⁴ Similar reservations must apply to Hellman and Thompson's claim that "incommensurable theories, between which no satisfactory translation is possible, could in principle be equivalent in the sense of co-determining one another".¹⁵

To summarise the picture so far, we can identify two very broad lines of approach to the problems of conceptual change, commensurability and progress. The first focuses on semantical concepts and attempts to locate a stability or continuity of reference or meaning in the development of scientific theories. On the whole, it tends to downgrade and underplay

¹² See CHURCHLAND (1979: Chap. 3). According to Churchland, a reduction need not embody a (meaning-preserving) translation between the theories concerned, nor is it required that the conceptual framework of the reducing theory can accommodate an accurate representation of the reduced theory. Like Kuhn, he holds that in many cases (as in the transition from classical to relativistic mechanics) a supplanting theory can at best be logically related to some modified and approximated version of its predecessor. However, in speaking of such cases as instances of reduction, it appears that Churchland has deprived the concept of reduction of just those features needed to ensure that theory-replacement via reduction is an empirically well-grounded and rational process.

¹³ See PEARCE (1982a,b, 1987).

¹⁴ This is (roughly) what I call Stegmüller's *rationality thesis*, since it carries the implication (explicitly formulated by STEGMÜLLER (1975)) that scientific progress can be rationally appraised (via the concept of reduction) even in cases where a new and revolutionary theory is incommensurable with its predecessor. Obviously, the *category* of appraisal here (reduction) is a familiar one, though its interpretation is non-standard.

¹⁵ HELLMAN and THOMPSON (1977: p. 334). I shall not present a detailed rebuttal of this claim here, but merely remark that it seems to be open to many of the same objections that can be raised against Stegmüller's thesis; see PEARCE (1985, 1987).

conceptual change, doing a disservice to the historical process of language creation in science.¹⁶ At the same time it fails in its principal task of showing that rival or successive theories in science are generally commensurable, or logically related by translation. The second approach relinquishes the idea of translatability or commensurability, and looks for alternative methods of cognitive comparison. However, the surrogate methods are either too weak to achieve their desired results, or else they turn out to be strong enough to entail translatability or commensurability of just the sort they were designed to avoid. In neither case have we advanced very far on the road to taking conceptual change seriously whilst preserving the rational picture of scientific progress.

2. Commensurability formalised?

It is a recurrent feature in the history of ideas that the radicalism of one generation becomes the common wisdom of the next. However, I doubt that today, as Kuhnian revolutions are being discovered in many branches of knowledge almost as a matter of routine, that we are actually embracing as radical a view of science as the one Kuhn offered us 25 years ago. In digesting Kuhn's theory, it is plain that we have also been toning it down and making it more respectable. Even so, one cannot help viewing some recent developments in the philosophy of science with a certain irony. In the 1960s Kuhn's and Feyerabend's concept of incommensurability was a major weapon in their frontal attack on the "received view" of scientific theories, with its emphasis on logical methods and rational reconstructions. Now, in the 1980s, "incommensurability" has become such a familiar idiom of the philosophical vocabulary, that it has, in its turn, become the subject of logical explication, much like any other concept (explanation or reduction or confirmation) belonging to the received view of theories.

Despite the apparent incongruity of treating commensurability as something to be formally defined, it is worth considering briefly one proposed explication to see whether it might in fact help to clarify Kuhn's thesis.

¹⁶ I mean to indicate a *tendency* here, rather than a general rule. In fact, there have been some valuable attempts to analyse continuity of reference in a context of conceptual change, notably by PRZEŁĘCKI (1979, 1980). Przełęcki does not, however, aim to establish (or deduce) linguistic commensurability or translatability. Non-translatability is rather an explicit feature of his analysis.

The proposal I shall examine has originated from BALZER (1985), though it has also been discussed and further refined by STEGMÜLLER (1986).¹⁷

Suppose we deal with two theories T and T' that are reconstructed by specifying their respective classes of models M and M' , the elements of these classes being semantical structures for languages L and L' in that order. Imagine further that we can define a class of T' -models each member of which can be correlated uniquely with some model of T . In other words, there is a partial function, F , say, mapping T' -models into T -models. In structuralist terms such a function satisfies the principal condition for determining a reduction of T to T' , and this is actually the situation in which Balzer's criterion of commensurability is supposed to apply. Leaving aside the matter of reducibility, we can look simply at the question: What properties should a syntactic translation Γ of L into L' possess in order to match or be compatible with the given semantic correlation F ? The obvious answer is that if Γ maps L -sentences into L' -sentences, then for any model \mathfrak{M} in the domain of F and any L -sentence θ , we require

$$F(\mathfrak{M}) \models \theta \quad \text{iff} \quad \mathfrak{M} \models \Gamma(\theta). \quad (1)$$

When this condition holds we can say that Γ respects F or that Γ is a companion of F .

Now, Balzer and Stegmüller argue that for T and T' to be commensurable, a function F of the above sort must exist and must possess a companion translation Γ . However, some further conditions must also be fulfilled; among them are (i) that L and L' share some common predicates \mathbf{R} ; (ii) that for any shared predicate $R \in \mathbf{R}$, there is some model $\mathfrak{M} \in \text{Dom}(F)$, such that $\mathfrak{M}(R) = F\mathfrak{M}(R)$; (iii) that for any L -sentence θ containing only shared predicates from \mathbf{R} , $\Gamma(\theta) = \theta$. In other words, it is required that Γ be a homographic or *literal* translation, in the sense that sentences common to the languages of T and T' count as their own translations (condition (iii)). Moreover, for any shared predicate R , at least one T -model assigns the same extension to R as does one of its counterpart models in T' (condition (ii)).

We have here a mixture of syntactic and semantic constraints that are supposed to establish commensurability by ensuring that the extensional

¹⁷ More recent attempts to improve on this explication are given by BALZER *et al.* (1987) and by SCHROEDER-HEISTER and SCHAEFER (1989). For reasons of space, I cannot discuss them here; see, however, PEARCE (1989).

or referential aspects of meaning are preserved under a suitable translation.¹⁸ In addition, this explication is intended to protect Stegmüller's thesis that reduction does not imply commensurability, because in cases where F represents a reduction relation, the fact that it may have a companion translation is no proof that the extra conditions for commensurability will be satisfied.

A moment's reflection, however, reveals that these constraints (i)–(iii) cannot be the appropriate ones. For instance, if T and T' are mutually inconsistent in the sense that for some sentence entailed by the one theory, its negation is entailed by the other, then, under this explication, the two theories cannot be commensurable. Hence, according to this account, any two theories with conflicting observational consequences will have to be counted as incommensurable. Moreover, the Balzer–Stegmüller approach appears to confuse the literalness of translation with its preservation of semantic values like reference. Literalness is an appropriate property of translation whenever shared expressions possess the same meaning in either language or theory. But this kind of meaning-invariance should not be assumed *prior to* translation. If it occurs, then it is a property that an adequate translation ought to reveal. If there is a variance of meaning between T and T' , then some expression θ of L , say, will have a different meaning in L' and consequently a translation that re-expresses θ in the L' -context will not be an identity mapping. In short, to demand of an adequate translation that it be literal is tantamount to reducing incommensurability to mere variance of meaning.

3. Kuhn on incommensurability

Kuhn himself has returned recently to the theme of incommensurability, attempting to buttress his thesis from the many onslaughts it has met over the years. He now treats as secondary the pragmatic and methodological aspects of incommensurability, regarding it primarily as a logical or semantical thesis to be defended in part by considerations from the philosophy of language.¹⁹ As in *SSR*, he continues to hold that the languages of theories like classical mechanics and phlogiston chemistry are untranslatable into the conceptual schemes that replaced them: those of special relativity and of oxygen chemistry. He also attempts to provide

¹⁸ See STEGMÜLLER (1986: Chap. 10).

¹⁹ See KUHN (1983a,b).

a more detailed analysis of those features of meaning that are in general supposed to be invariant under translation.

Kuhn's case that meaning-variance, conceptual change and conceptual growth are part and parcel of the scientific enterprise is a strong one. However, his new analysis does not strengthen one bit his arguments for non-translatability in specific cases. He repeats his former claim that Newtonian concepts like mass and force are untranslatable into the language of modern physics, adding the new argument that such concepts can only be *learnt* in the Newtonian context where the Second Law of Motion applies. Yet the law is not a tautology or analytic statement: it can even, he now holds, be falsified. It is however, "necessary" in the sense that if the law fails the Newtonian concepts would, in his words, be "shown not to refer".²⁰

This is perplexing. That the Second Law can *not* be falsified was one of the major theses of *SSR* which even many of Kuhn's critics accepted and tried to elucidate. That concepts like Newtonian mass and force might fail to refer, or lack empirical denotation, is an idea that Kuhn's principal opponents, the logical empiricists, had defended at least since Philipp Frank's *Foundations of Physics*. Moreover, having nicely distinguished the process of acquiring a new language from that of translating a familiar one into a new framework, Kuhn astonishingly proceeds to conflate learning with translation by applying properties of the former to argue against the possibility of the latter.

Much of Kuhn's recent discussion of commensurability simply deviates too far from the central issues. He continues to dwell on the question whether historical *texts* can be translated into modern language, as if the logical problem of intertheoretic comparisons were a matter to be resolved purely by textual analysis instead of by rationally reconstructing theories. Texts may be a useful and even indispensable basis for interpretation and rational reconstruction, but the real question at issue is whether a theory or conceptual scheme so reconstructed admits of translation into a later framework. Moreover, in assessing the rationality of conceptual and theoretical changes in science, it is reasonable to assume that at the moment when scientists choose in favour of one theory and against a rival, perhaps older, theory, both of the languages in question are understood. It is therefore scarcely appropriate, as Kuhn and others have imagined, to liken this situation to that of radical translation in Quine's sense.

²⁰ KUHN (1983b: p. 567).

Kuhn now analyses linguistic meaning in terms of the taxonomic categories employed by users of a language and the similarities and differences between terms in the way they apply to nature. In his view, meanings are determined, not by the individual criteria or rules of application employed by the language user, but rather by the global structure of such rules. It is this structure, of taxonomy and of similarities and differences between terms, which an adequate translation must, according to Kuhn, preserve. The upshot is that translation is only possible between languages possessing structurally identical taxonomic categories which are similarly interrelated.

Kuhn's discussion here is too vague to attempt a more detailed reconstruction. However, even at a superficial level of analysis, structural identity of taxonomy must strike one as an excessively strong requirement to impose on interlinguistic translation. It might be the kind of property one would expect to find in the case of conceptual frameworks that are fully equivalent, or intensionally isomorphic to one another. But even under such a holistic conception of meanings as Kuhn adopts (whereby alterations in some meanings inevitably ring changes in the conceptual scheme as a whole), it is hard to sustain the idea that translation, when it exists, must be both complete and "full". On the contrary, since the conceptual schemes of science are typically "open-ended" in that intentions are not fully determined by theories, it is natural to conceive that the translation process may be both "incomplete", in the sense that it may transform some but not all features or expressions of one language into another, and "partial" in the sense that the conceptual categories of the source framework become embedded into those of the target framework, without the correlation necessarily determining a full equivalence. In fact, taken at face value, Kuhn's formulation of commensurability should permit both these types of "incompleteness". On the one hand, he speaks of "partial" and of "local" incommensurability. On the other hand, he regards two theories to be incommensurable "if there is no language, neutral or otherwise, into which both theories, conceived as sets of sentences, can be translated without residue or loss".²¹ Nothing here precludes the possibility that translation might be a one-way process, that a scientific conceptual framework, or some extension of it, might sustain a translation of its predecessor, without itself being translatable back into the latter.

Some recent efforts to provide a more detailed analysis of translation

²¹ KUHN (1983a: p. 670).

between conceptual frameworks also seem to favour the more “liberalised” approach. For instance, Haim Gaifman’s theory of ontology and conceptual frameworks allows both for differing grades of “openness” in a framework and differing degrees of ontological “overlap” between frameworks.²² Moreover, Gaifman argues persuasively that in at least some prominent cases where rival conceptual schemes are by no means fully equivalent, nevertheless partial translations can be found which establish a significant “common ground” of ontology and meaning.

4. Correspondence, commensurability and progress

Gaifman’s approach is a promising one in cases where we are trying to identify some stable, common, even neutral ground shared by rival conceptual schemes and scientific world-views. However, the history of science, and especially of physics, also exhibits examples of theoretical change where it is not clear that there is anything like a “neutral” perspective from which we can make cognitive comparisons. One circumstance I have in mind is where a theory contains its predecessor as a limiting case. Classical and special relativistic mechanics is perhaps the best-known and most controversial example. Many philosophers would like to argue, against Kuhn and Feyerabend, that we have here a clear example of reduction; Newton’s theory being reducible to its successor. However, we also have all the ingredients that make reduction in the standard, deductive-explanatory sense problematic: the *prima facie* syntactic incompatibility of the two theories, the fact that approximating and even counterfactual assumptions may be needed to express logical relations between them, the apparently radical changes of conceptual scheme occurring in the transition from classical to relativistic physics, throwing translatability and commensurability into doubt.

After Niels Bohr first coined the phrase *Korrespondenzprinzip* to express the idea that such limiting-case relations, as that of classical to special relativistic mechanics (CM and RM), and of classical to quantum mechanics, embody a fundamental principle of scientific development, philosophers of science have analysed the correspondence relation in detail. The way was led by the Soviet writer I.V. Kuznietsov in his book *The Correspondence Principle in Contemporary Physics and its Philosophical Meaning*, published in 1948. Following Kuznietsov, the

²² See GAIFMAN (1975, 1976, 1984).

correspondence principle has become not only an important instrument for Soviet and other Marxist philosophers studying the dialectics of scientific progress, it has also become a central topic of discussion by philosophers of science of all persuasions.²³

Today, there are many, widely-differing interpretations of the correspondence relation, as applied say to classical and relativistic mechanics. They serve to illustrate quite distinct approaches to the problem of describing conceptual and theoretical change in science. The structuralist analysis (as presented by STEGMÜLLER, 1986) exemplifies a view which takes conceptual change seriously, but which grasps this change exclusively at the structural or model-theoretic level; matters of language and translation being suppressed. CM and RM are independently reconstructed and held to be related by an *approximate* reduction. This differs from exact reduction, firstly in that to represent approximation processes a topological structure is imposed on the classes of the theories' models, and secondly in that the structural correlation connects models of classical mechanics with "relativistic" structures that are strictly speaking no longer *models* of RM.²⁴

A somewhat different, though related, account has been developed in West Germany by physicists and philosophers of physics, such as Günther Ludwig, Erhard Scheibe and Jürgen Ehlers.²⁵ It shares the idea that approximate relations between physical theories can be studied by assigning so-called *uniform* topologies to their spaces of models. But in Ehlers' treatment, for example, there is no attempt to separate the classical from the relativistic conceptual framework or to provide independently motivated axiomatisations. One works in a shared mathematical framework of differential manifolds and tensor fields, and the main idea is to represent the difference between CM and RM as consisting of the different values they assign to a single, shared physical constant (e.g. maximum signal speed, space-time curvature). But the result is a very elegant mathematical treatment of the limiting-case relation, and of certain key problems related to scientific progress. The perspective taken is very much that of "modern physics" with its reformulation of classical theories in conformity with a later standpoint. It is, in short, closer to the stance of the

²³ For a survey, especially of Soviet and Polish writings on the correspondence principle, see KRAJEWSKI (1977).

²⁴ For a general analysis and explication of approximate reduction, see MOULINES (1980), MAYR (1981) and BALZER *et al.* (1987). The case of CM and RM is treated by STEGMÜLLER (1986: Chap. 8).

²⁵ See especially LUDWIG (1978, 1981), SCHEIBE (1973), EHLERS (1986).

textbook author who, as Kuhn would say, rewrites history backwards, rather than that of the historian or philosopher concerned with the actual transition from one scientific paradigm to another.²⁶

In Kuhn's own account of the relation of classical to relativistic mechanics, limiting-case correspondence does not establish conceptual links between the two paradigms. For Kuhn, therefore, the classical limits of relativistic theories are not classical "theories" at all, but rather reformulated and approximated versions of them. There is no translation of Newtonian concepts into the language of RM. In short, the three interpretations I have mentioned seem to offer us a choice between (i) an ahistorical account which largely suppresses conceptual change (Ehlers), (ii) an exclusively structural analysis, which avoids language (Stegmüller), or (iii) conceptual change without commensurability (Kuhn). There is, however, a fourth interpretation that offers both a precise characterisation of the limiting-case relation *and* an analysis of conceptual change. This approach was initiated by Veikko Rantala and was later further developed in our joint papers on theory structure and intertheory relations.²⁷ I want to suggest now that it might provide an adequate model of rational theory change which could help to resolve some of the problems that still divide Kuhn from his critics.

Rantala's and my conception of theories is a structural one; but it differs in several significant ways from the *structuralist* view of Sneed and Stegmüller as well as from the so-called *semantic* view adopted by van Fraassen, Giere and others.²⁸ One of the chief differences concerns our use of *logic*. In contemporary philosophy of science there has been much discussion about the proper role and function of logic as an instrument of metascientific and methodological analysis. A good deal of this discussion has been severely *critical* of the way logic was traditionally used in positivist and empiricist methodologies of science and, as part of this antipositivist backlash, the "semantic" and "structuralist" approaches

²⁶ EHLERS (1986) treats in some detail two important examples of the limiting-case relation: the theory pairs (Galilei-invariant, Lorentz-invariant collision mechanics) and (Newtonian, general relativistic gravitation theory). For these examples, Ehlers affirms: "I am not concerned with the histories of the theories in question or with the intuitive ideas, motivations, manner of presentation etc. of their originators, but with the rational reconstructions of those theories, logical relations between them, and with their relations to experience." (1986: p. 387).

²⁷ See e.g. RANTALA (1979), PEARCE and RANTALA (1983a, 1984a, 1984b, 1985). Besides the four lines of approach mentioned here, various other different analyses of the correspondence relation have been offered; see e.g. POST (1971), KRAJEWSKI (1977).

²⁸ e.g. VAN FRAASSEN (1970, 1980, 1986), GIERE (1983).

advocate mathematical rather than logical or metamathematical methods of analysis. This criticism is well-founded to the extent that it is directed towards certain of the highly idealised logical assumptions often found in standard models of theory structure and testing, and in popular accounts of explanation, reduction, confirmation etc. In many cases, however, this criticism is misplaced in that it labours under a rather prevalent misconception among philosophers that logic is *first-order* logic. However, there is no more truth to the claim that logic is first-order logic than there is to the claim that scientific reasoning is limited to the resources of first-order logic. The fact that some logical models of science have been static, restrictive and heavily idealised is no ground for thinking that the best “logic” of science is no logic whatsoever.

Over the past two or three decades, many of the most important advances within logic itself have been aimed at bringing it closer to the actual practice of mathematics, science and natural language. Philosophers of science have been slow to exploit these developments, with the consequence that a gulf has opened up between logical and historical approaches to science; it being a widespread fallacy that formal methods and logical reconstructions must do violence to the *real* history and practice of science. One way to correct this misleading image is to think of logic as a flexible and open-ended tool of rational reconstruction, rather than as straitjacket into which everything can be fitted come what may. The structuralist and the semantic approaches have already taught us that “formal” is not a synonym for “unrealistic”; but in trying to study structures or models whilst ignoring the model *theory*, in the logical sense, they act like the prisoner who evades the straitjacket by accepting to be handcuffed.

In the account of theory structure and development that Rantala and I propose, it is not the scientific theory that has to fit the constraints of logic, but rather the other way round: logic and semantics are variables, contoured to fit the theory or the metatheory at hand. Generality can be combined with uniformity by adopting the perspective of *abstract* model theory which studies properties not of this-or-that logic, but of logics in general or logics of a certain type. This means that one can continue to represent a scientific theory by a suitable class of models, make explicit that a certain vocabulary or set of non-logical terms is associated with the theory, whilst leaving open the question whether a particular logic or language (syntax + semantics) is “the right one”. Without committing ourselves in advance to a particular choice of logic, first-order or otherwise, we can still define intertheoretic translation, explanation and reduc-

tion in syntactic as well as semantic terms. Without going into details, let me try to convey the flavour of this method when applied to a particular case of conceptual change, that of classical and relativistic mechanics.²⁹

Consider two very simple versions of particle mechanics, one satisfying Newton's 2nd Law, the other satisfying Minkowski's force law. The two theories can be taken to share a common stock of primitive terms, in particular terms designating the spatial location of particles, their mass, the forces acting on them, etc. In addition, RM contains a new primitive, c , denoting the velocity of light. The models of both theories can be based on some underlying model of real analysis, and can be assumed to comprise finite sets of particles and a set of time points. Let these classes of models be M and M' , respectively. Now, the fact that Minkowski's Law approaches the 2nd Law in the limit of low particle velocities v , i.e. $v/c \rightarrow 0$, means that some models of CM are "very close" to models of RM. As Rantala first showed in 1977, this idea of closeness can be made exact using the tools of nonstandard analysis. Formally, any standard model of CM³⁰ in which velocities and accelerations of particles are bounded (by standard real numbers) can be canonically represented as the limit or "standard approximation" of a suitable relativistic model whose underlying model of analysis is nonstandard and whose particle velocities are infinitesimal compared with the speed of light. Graphically, one can define a function F mapping a definable subclass K' of M' into M .

This semantic correspondence F has three remarkable features: it performs a simple mathematical construction collapsing nonstandard entities in a model to their (uniquely obtainable) standard parts;³¹ it converts a relativistic model into an infinitesimally close *classical* model; and it has a companion translation Γ . In fact, Γ can be defined as a recursive and non-homographic mapping of $L(\tau)$ -formulae into $L(\tau')$ -formulae, where L is a suitable logic and τ , τ' denote, respectively, the vocabularies of CM and RM.³² We obtain, once again, schema (1), viz.

²⁹ This example, described only very sketchily below, is reconstructed at length by PEARCE and RANTALA (1984a), to which the reader is referred for further details. A discussion of the implications of this case-study for the problems of meaning-change and incommensurability is given by PEARCE (1987: Chap. 7).

³⁰ That is to say, a model whose underlying model of analysis is standard.

³¹ The "standard part" relation is usually defined for real numbers, but it can be generalised to apply to arbitrary functions, relations, etc. See PEARCE and RANTALA (1984a).

³² L can be taken to be a strong infinitary logic in which the relevant model classes, M , M' , K' as well as the range, say K , of F are definable. Thus, we associate to CM the language $L(\tau)$ and to RM the language $L(\tau')$. Since we have supposed the basic vocabulary of RM to be that of CM plus an additional constant c , we have $\tau' = \tau \cup \{c\}$.

for all models $\mathfrak{M} \in K'$ and all $L(\tau)$ -sentences θ ,

$$\mathfrak{M} \models_L \Gamma(\theta) \text{ iff } F(\mathfrak{M}) \models_L \theta.^{33}$$

Since Γ is not a literal translation, it expresses the fact that there is indeed a conceptual change between CM and RM. An atomic formula involving, say, classical “mass” is transformed into a complex formula of RM involving extra terms. However, relative to the semantic correlation F between classical and relativistic models, we can say that Γ preserves, in a suitable sense, reference. (Notice that classical mass is not being identified here with either rest mass or relativistic mass.) Moreover, using this translation one can express a deductive relation between the two theories. If θ, θ' are, respectively, suitable formalisations of the 2nd Law and the Minkowski Law (in $L(\tau)$ and $L(\tau')$, respectively), we obtain:

$$\theta', \beta \models_L \Gamma(\theta), \tag{2}$$

where β formalises the limit condition $v/c \approx 0$.

Likewise, we can represent through this translation classical problem solutions in the relativistic framework. Assume that the range K of F is axiomatised by some $L(\tau)$ -sentence, α . Then those classical problem solutions or explanations that are captured by a schema of the form

$$\alpha, \Phi \models_L \sigma, \tag{3}$$

where Φ denotes the initial and boundary conditions, have an approximate validity in the RM framework, by virtue of

$$\theta', \beta, \Gamma(\Phi) \models_L \Gamma(\sigma). \tag{4}$$

Schema (2) here differs from the usual account in that the left-hand side, the explanans, is a consistent extension of RM. It requires us to assume neither that c is infinite nor that particle velocities are zero. Secondly, on the right-hand side, instead of an “approximated” or “corrected” version of CM, we have a *translation* of the central Newtonian Law into the relativistic framework. Explicating the correspondence relation in this way, one is able to take account of conceptual change whilst interpreting the framework of RM as a language rich enough to sustain a translation of Newtonian concepts and laws, and to provide a means of rationally appraising CM. Moreover, RM can be seen to explain (approximately) the successful part of CM under the limit condition β .

³³ \models_L here denotes the L -consequence relation.

Here nonstandard analysis provides a method of replacing what would normally be regarded as subjunctive, counterfactual conditions of explanation by equivalent, indicative, truth-functional conditions. However, as Rantala has recently shown, this translation can also be used to support a "counterfactual" explanation of CM by RM, taking counterfactual conditionals to be interpreted in the Lewis-style semantics of possible worlds.³⁴

The general picture, then, is of meaning change and conceptual growth, but with the newer theory being capable of acting as a vehicle of rational comparison and appraisal. One can therefore agree at this point with Kuhn that there is no "theory-neutral" perspective: translation is a one-way process, proceeding from the vantage point of the newer conceptual framework. But because translation and semantic correspondence do exist, there is nothing mysterious or irrational about the process of cognitive comparison. Rational preferences can be based on empirical grounds, just as physicists would wish to maintain.

The model I have described still has an air of being superficial and idealised; but it can be generalised and extended in several different directions. The method of nonstandard analysis is actually more general than its name suggests, and can be applied to qualitative as well as different sorts of quantitative theories. The abstract model-theoretic framework can be used not only to represent other types of inter-theoretic relations but also other features of conceptual growth not directly related to models or laws.³⁵

The picture I have been sketching can also be usefully combined with various different models of scientific progress. Since it makes translation an explicit part of our analysis of theoretical change, it offers the means to relocate the questions, empirical problems and solutions of one

³⁴ See RANTALA (1986, 1987).

³⁵ For a discussion of reduction and other intertheoretic relations in this framework, see PEARCE and RANTALA (1983c) and PEARCE (1987). An important feature of conceptual change in physics that does not directly involve laws concerns the manner in which the *symmetries* or *invariances* associated with a theory are preserved or transformed when that theory is replaced. The matter of symmetry change is dealt with by PEARCE and RANTALA (1983b, 1984a). The methods of non-standard analysis discussed here in connection with the correspondence relation are of course largely independent of the specific abstract model-theoretic conception of theories that Rantala and I have developed. Thus, nonstandard analysis can in principle be applied within other metascientific frameworks, providing that they do not (like the structuralist and semantic approaches) explicitly eschew logical concepts and techniques. However, the perspective of abstract model theory is particularly well-suited to the present type of analysis, in that (i) it brings out clearly the way in which logics of varying types and strengths may be applied in metascientific studies, and (ii) it allows for a more differentiated discussion of the implications that such studies may have for philosophical questions concerned with meaning-change, rationality, progress, etc.

research tradition within the conceptual schemes of its rivals. It may therefore enhance our grasp of scientific progress, when progress is viewed, with Laudan, as a matter of comparative problem-solving effectiveness, or, with Hintikka, in terms of an interrogative or question-answering model of scientific inquiry. Likewise, it could be used to supplement critical realist accounts of scientific growth. For instance, Niiniluoto's theory of truthlikeness offers exact measures of the distances between competing laws and problem-solutions *within* any given conceptual scheme. Wherever rival theories employ characteristically different conceptualisations of a problem, translation will be needed to compare these measures; and in those cases like the one mentioned here, where no fully independent or neutral system of representation is available, the natural strategy is to look at distances and degrees of truthlikeness as they appear under translation into a suitable "target" framework.

5. Concluding remarks

Thus, in reconstructing scientific growth I think we need not assume either that translation is unnecessary because science possesses something like a universal language, or that translation is somehow "already achieved". Nor do we have to accept Kuhn's thesis that translation is often impossible. We can and we should represent scientific change by explicitly building translation and conceptual growth into our models of progress. I am not claiming that translation will always be trivial, nor even that it will always exist. Where rival theories in science share a common experimental basis, there are natural ways to find the sorts of semantic correspondences on the basis of which translations can be constructed. Where a marked shift in the experimental background accompanies the transition to a new theory, the logical connections between the old and the new world-views will be harder to establish. Whether such shifts can be so disruptive as to preclude altogether translation and logical comparison, remains an open question. However, radical changes of experimental practice do sometimes take place, as is illustrated, today for example, by the developments that have occurred in high energy physics since the first appearance of Kuhn's book. Finding an adequate conceptual reconstruction of the transition from the early quark models of matter to the later paradigm of high energy physics, based on quantum chromodynamics and the quark-gauge theories of matter, remains an important challenge for both historians and logicians of science in the future.

References

- BALZER, W., 1985, *Incommensurability, reduction and translation*, *Erkenntnis* 23, pp. 255–267.
- BALZER, W., MOULINES, C.U. and SNEED, J.D., 1987, *An Architectonic for Science* (D. Reidel, Dordrecht).
- BALZER, W., PEARCE, D. and SCHMIDT, H.-J., eds., 1984, *Reduction in Science: Structure, Examples, Philosophical Problems* (D. Reidel, Dordrecht).
- CHURCHLAND, P.M., 1979, *Scientific Realism and the Plasticity of Mind* (Cambridge University Press, Cambridge, UK).
- EHLERS, J., 1986, *On limit relations between, and approximative explanations of, physical theories*, in: R. Barcan Marcus *et al.*, eds., *Logic, Methodology and Philosophy of Science VII* (Elsevier, Amsterdam).
- FIELD, H.H., 1973, *Theory change and the indeterminacy of reference*, *J. Philosophy* 70, pp. 462–481.
- FRAASSEN, B. VAN, 1970, *On the extension of Beth's semantics of physical theories*, *Phil. Sci.* 37, pp. 325–339.
- FRAASSEN, B. VAN, 1980, *The Scientific Image* (Clarendon Press, Oxford).
- FRAASSEN, B. VAN, 1986, *Aim and structure of scientific theories*, in: Barcan Marcus *et al.*, eds., *Logic, Methodology and Philosophy of Science VII* (Elsevier, Amsterdam).
- GAIFMAN, H., 1975, *Ontology and conceptual frameworks*, Part I, *Erkenntnis* 9, pp. 329–335; 1976, Part II, *Erkenntnis* 10, pp. 21–85.
- GAIFMAN, H., 1984, *Why language?*, in: Balzer *et al.*, eds., *Reduction in Science: Structure, Examples, Philosophical Problems* (D. Reidel, Dordrecht).
- GIERE, R., 1983, *Testing empirical hypotheses*, in: J. Earman, ed., *Testing Scientific Theories*, *Minnesota Stud. Phil. Sci.* Vol 10 (University of Minnesota Press, Minneapolis).
- HARTKAEMPER, A. and SCHMIDT, H.-J., eds, J. 1981, *Structure and Approximation in Physical Theories* (Plenum Press, New York).
- HELLMAN, G. and THOMPSON, F., 1975, *Physicalism: ontology, determination and reduction*, *Journal of Philosophy* 72, pp. 551–564.
- HELLMAN, G. and THOMPSON, F., 1977, *Physicalist materialism*, *Noûs* 11, pp. 309–345.
- KRAJEWSKI, W., 1977, *Correspondence Principle and the Growth of Science* (D. Reidel, Dordrecht).
- KUHN, T.S., 1983a, *Commensurability, comparability, communicability*, in: P. Asquith and T. Nickles, eds., *PSA 1982* (Philosophy of Science Assoc., East Lansing).
- KUHN, T.S., 1983b, *Rationality and theory choice*, *J. Philosophy* 80, pp. 563–570.
- LAUDAN, L., 1977, *Progress and its Problems* (University of California Press, Berkeley, Los Angeles, London).
- LUDWIG, G., 1978, *Die Grundstrukturen einer physikalischen Theorie* (Springer-Verlag, Berlin).
- LUDWIG, G., 1981, *Imprecision in physics*, in: A. Hartkaemper and H.-J. Schmidt, eds., *Structure and Approximation in Physical Theories* (Plenum Press, New York).
- MAYR, D., 1981, *Investigations of the concept of reduction*, II, *Erkenntnis* 16, pp. 109–129.
- MOULINES, C.-U., 1980, *Intertheoretic approximation: The Kepler–Newton case*, *Synthese* 45, pp. 387–412.
- NIINILUOTO, I. and TUOMELA, R., eds., 1979, *The Logic and Epistemology of Scientific Change* (*Acta Philosophica Fennica* 30), Amsterdam.
- PEARCE, D., 1982a, *Stegmüller on Kuhn and incommensurability*, *Brit. J. Phil. Sci.* 33, pp. 389–396.
- PEARCE, D., 1982b, *Logical properties of the structuralist concept of reduction*, *Erkenntnis* 18, pp. 307–333.

- PEARCE, D., 1985, *Remarks on physicalism and reductionism*, in: G. Holmstroem and A. Jones, eds., *Action, Logic and Social Theory* (Acta Philosophica Fennica 38), Helsinki.
- PEARCE, D., 1987, *Roads to Commensurability* (D. Reidel, Dordrecht).
- PEARCE, D., 1989, *Translation, reduction and commensurability: a note on Schroeder-Heister and Schaefer*, *Phil. Sci.* 56 (1).
- PEARCE, D. and RANTALA, V., 1983a, *New foundations for metascience*, *Synthese* 56, pp. 1–26.
- PEARCE, D. and RANTALA, V., 1983b, *The logical study of symmetries in scientific change*, in: P. Weingartner and H. Czermak, eds., *Epistemology and Philosophy of Science* (Hoelder-Pichler-Tempsky, Vienna).
- PEARCE, D. and RANTALA, V., 1983c, *Logical aspects of scientific reduction*, in: P. Weingartner and H. Czermak, eds., *Epistemology and Philosophy of Science* (Hoelder-Pichler-Tempsky, Vienna).
- PEARCE, D. and RANTALA, V., 1984a, *A logical study of the correspondence relation*, *J. Philos. Logic* 13, pp. 47–84.
- PEARCE, D. and RANTALA, V., 1984b, *Limiting case correspondence between physical theories*, in: Balzer *et al.*, eds., *Reduction in Science: Structure, Examples, Philosophical Problems* (D. Reidel, Dordrecht).
- PEARCE, D. and RANTALA, V., 1985, *Approximative explanation is deductive-nomological*, *Philosophy of Science* 52, pp. 126–140.
- POST, H., 1971, *Correspondence, Invariance and Heuristics*, *Stud. Hist. Phil. Sci.* 2, pp. 213–255.
- PRZEŹECKI, M., 1979, *Commensurable referents of incommensurable terms*, in: I. Niiniluoto and R. Tuomela, eds., *The Logic and Epistemology of Scientific Change* (Acta Philosophica Fennica 30), Amsterdam.
- PRZEŹECKI, M., 1980, *Conceptual continuity through theory changes*, in: R. Hilpinen, ed., *Rationality in Science* (D. Reidel, Dordrecht).
- PUTNAM, H., 1973, *Explanation and reference*, in: G. Pearce and P. Maynard, eds., *Conceptual Change* (D. Reidel, Dordrecht).
- PUTNAM, H., 1977, *Meaning and reference*, in: S. Schwartz, ed., *Naming, Necessity and Natural Kinds* (Cornell Univ. Press, Ithaca).
- RANTALA, V., 1979, *Correspondence and non-standard models: a case study*, in: I. Niiniluoto and R. Tuomela, eds., *The Logic and Epistemology of Scientific Change* (Acta Philosophica Fennica), Amsterdam.
- RANTALA, V., 1986, *Counterfactual reduction* (to appear in K. Gavroglu *et al.*, eds., *Proc. Int. Conf.: Criticism and the Growth of Knowledge: 20 Years After* (Thessaloniki)).
- RANTALA, V., 1987, *Explaining superseded laws*, Abstracts of the 8th Int. Congress of Logic, Methodology & Philosophy of Science, Moscow.
- SCHEIBE, E., 1973, *The approximative explanation and the development of physics*, in: P. Suppes *et al.*, eds., *Logic, Methodology and Philosophy of Science, IV* (North-Holland, Amsterdam).
- SCHROEDER-HEISTER, P. and SCHAEFER, F., 1989, *Reduction, representation and commensurability of theories*, *Phil. Sci.* 56 (1).
- SHAPERE, D., 1982, *Reason, reference, and the quest for knowledge*, *Phil. Sci.* 49, pp. 1–23.
- STEGMÜLLER, W., 1975, *Structures and dynamics of theories: some reflections on J.D. Sneed and T.S. Kuhn*, *Erkenntnis* 9, pp. 75–100.
- STEGMÜLLER, W., 1986, *Theorie und Erfahrung, Dritter Teilband* (Springer-Verlag, Berlin).

This Page Intentionally Left Blank

SCIENTIFIC METHOD AND THE OBJECTIVITY OF EPISTEMIC VALUE JUDGMENTS

KRISTIN SHRADER-FRECHETTE

University of South Florida, Tampa, FL 33620-5550, USA

Introduction

Current philosophical fashion, as exhibited through DAVIDSON (1984a: p. 194; 1984b: p. 425) and RORTY (1986a,b) and earlier through Wittgenstein and Dewey, has it that there is no such thing as the “best explanation” in science or anything else; there is just the explanation that best suits some particular explainer. If this is correct, then we ought to give up doing philosophy of science and turn to sociology and psychology.

Without taking a position of the realism issue, on the demarcation problem, on whether science is a natural kind, and on whether scientific rationality is different from other kinds of rationality, I shall argue that we can talk about “best explanations”. To do so, however, we need to understand a whole spectrum of views on the question of whether there are any characteristics of science that are immune to change and hence that guarantee scientific rationality, a necessary condition for scientific objectivity (see PUTNAM 1981, SIEGEL 1985: p. 532).

At one end of the spectrum, the “pluralist end”, are epistemological anarchist FEYERABEND (1975: p. 177) and others who believe, as Feyerabend put it, that “no system of [scientific] rules and standards is ever safe” (1977: p. 379). At the other end of the spectrum, the “universalist end”, are logical empiricists, such as Carnap and Schlick, and some Popperians, all of whom believe that at least some criteria for theory choice are fixed (HEMPEL 1979: p. 55; CARNAP 1950, 1952, 1967; SCHLICK 1959: pp. 209–227; SCHEFFLER 1982: p. 9). In the middle, between the pluralists and the universalists, are philosophers like SHAPER (1984), LAUDAN (1984), and GIERE (1985), each of whom holds a slightly different version of naturalism (see POPPER 1965: p. 52; McMULLIN 1986: p. 10). Although, on other accounts, Shapere, Laudan and Giere might

not all be termed “naturalists”, at least one of several naturalistic beliefs shared by them is that, although there are no absolute, *a priori*, normative rules of scientific method, theory choice nevertheless can be rational. As Shapere puts it, for example, science is able to proceed rationally, in the light of its best beliefs, even though it is open to change in all respects, its subject matter, methods, standards and goals (SHAPER 1984: pp. 207ff., 350ff.).

I will defend a position, hierarchical naturalism, midway between the universalists and the naturalists. It is best defined in terms of four propositions: (1) following Hempel, there is at least one general, universal criterion or goal of theory or paradigm choice, viz., explanatory power as tested by prediction (HEMPEL 1979: p. 56; 1983: p. 91; SELLARS 1967: p. 410), even though (2) Hempel, Carnap, and others were wrong about the types of epistemic value judgments made in connection with this criterion. (3) Most of the remaining criteria for theory choice, although evaluated and interpreted in terms of how well they function as *means* to the *end* or goal of explanatory power, are both situation-specific or determined largely by practice. (4) If certain of these criteria or goals allow the possibility of intelligible debate and criticism by the scientific community, then they guarantee what I shall call “scientific objectivity”. Since I will not have time to discuss all four of these propositions, I will put off (for the time being) the problems associated with defending an internalist account of rationality (GOLDMAN 1980: p. 27), and instead rely mainly on the arguments of McMullin and Hempel for thesis (1). After arguing for claims (2)–(4), I will use an example from hydrogeology to illustrate my position.

1. Feyerabend, method and value judgments

According to Feyerabend, “there are no overriding rules which are adhered to under any circumstance; there is no ‘scientific methodology’” (1980: p. 61; 1978: p. 300) that is immune to change over time. A milder version of Feyerabend’s attack on universal rules of scientific method has been echoed among sociologists and philosophers of science who correctly recognize, as Scriven put it, that “science is essentially evaluative” (SCRIVEN 1980: p. 283).

Since value judgments, associated with the interpretation and application of scientific rules and goals, appear to be part of what allegedly renders these goals and rules unstable, we need to understand the degree

to which such judgments are purely subjective (see RUDNER 1980: p. 286). CAWS (1967: pp. 59–61) for example, attempts to rescue values from the realm of subjectivity by classifying both facts and values as species of facts. His account fails, however, in part because the many facets of values cannot be reduced to his one category of facts.

A better approach might be to specify the different types of values and the criteria for each type. SCRIVEN (1974) and McMULLIN (1983), for example, note that values can be either emotive, pragmatic, or cognitive; only emotive values have no place in science. Within cognitive value judgments, McMullin distinguishes *evaluating* from *valuing*. We can make a largely factual judgment and *evaluate* the extent to which a thing possesses a characteristic value, e.g., a theory possesses predictive power; or we can make a largely subjective judgment and *value* an alleged property, e.g., assess the extent to which a characteristic, such as simplicity, is really a value for a theory.

Like Hempel's "instrumental value judgments" and Scriven's "value-performance claims", what McMullin calls "evaluating" judgments assert that, if a specified goal or value is to be obtained, then a certain action is good. Like Hempel's "categorical judgments" and Scriven's "real-value claims", what McMullin calls "valuing" judgments state that a certain goal is *prima facie* good, independent of particular circumstances. Whereas post positivists such as McMullin and Scriven accept both instrumental and categorical value judgments, Hempel and others believe that the latter have no place in science, because they cannot be confirmed empirically (HEMPEL 1980: p. 263; 1979: pp. 45–66; 1983: pp. 73–100; SCRIVEN 1974; NAGEL 1961: p. 492).

2. Empirical confirmability and scientific objectivity

In making empirical confirmability a criterion for judgments in science, and thereby excluding categorical judgments of value, logical empiricists such as Carnap and Hempel err for at least two reasons. *First*, Aristotle claimed that wise persons realize the certainty characteristic of different kinds of judgments, and that they only demand a certainty appropriate to the type of investigation. The logical empiricists demand an inappropriate level of certainty because their requirement of confirmability would not allow scientists to decide on criteria for theory choice, since such choices could not be confirmed. Yet, as I shall illustrate later with an example

from hydrogeology, scientists do make judgments about weighting criteria for theory evaluation, and they do so all the time. This means not only that scientific practice does not sanction the logical empiricists' and some universalists' notion of the role of confirmability in science, but also that, if all judgments in science had to be confirmed, then science as we know it would come to a halt.

Second, even if all judgments in science could be empirically confirmed, once they were confirmed, then the judgments could never change. Then science would never progress, impotent theories would never be discarded, and scientific revolutions would never occur. However, science does progress, impotent theories often are discarded, and scientific revolutions do occur. Hence it is obvious both that all judgments in science are not confirmed, hence unchangeable, and indeed that they cannot be, if science is to improve as it has.

It is not reasonable to require empirical confirmability of all scientific judgments because it is not the only test of objectivity, either in science or anywhere else. For example, we often call a judgment "objective" if it is not obviously biased or subjective. Objectivity, in this sense, is not tied to certainty, as confirmability is, so much as it is linked to even-handed representation of the situation. What I call "scientific objectivity" (and I say "scientific" not because the objectivity is unique to science, but simply because it is characteristic of it) is closely related to this sense of objectivity, as even-handedness. Presumably one could be blamed for failure to be objective in this sense, if one were biased in a particular judgment. Since we do often blame people for not being objective, in a sense close to that about which I am speaking, it is clear either that objectivity in this sense must be attainable or that one can be more or less objective.

But how might one guarantee scientific objectivity in the sense of even-handed representation of the situation? It will not do to say that a judgment in science is objective if it fits the situation or "the facts", because (1) we do not want to beg the realism question, (2) we might not have all the facts, and (3) since every situation in science will be different, it is virtually impossible to specify, ahead of time, what an even-handed representation of the situation might be. If so, then we may not be able to come up with characteristics of scientific objectivity. Instead we may be able only to avoid the charge of violating scientific objectivity.

One way to avoid such a charge might be to subject our scientific judgments to review by the scientific community. If so, then judgments

might be said to possess scientific objectivity, in some minimal sense, if they can be subjected to criticism, debate, and amendment by the scientific community. On this account, although *scientific rationality* might be guaranteed by a scientist working *individually*, by pursuing a goal of explanatory power tested by prediction, *scientific objectivity* could only be guaranteed by scientists *collectively*. On this view, scientific rationality is a necessary condition for what I call scientific objectivity.

In addition, if epistemic or cognitive value judgments in science (e.g. theory A has more predictive power than theory B) were capable of being subjected to the criticism and evaluation of the scientific community, then even these value judgments could be said to possess scientific objectivity. This is not as implausible as it sounds, however, for several reasons. *First*, when I make an epistemic value judgment about two theories, for example, I am not talking merely autobiographically or subjectively. I am talking about two things having external referents and capable of being known and understood by other people. *Second*, the skills associated with making these judgments are a function of training, education, experience and intelligence. If so, objectivity does not require having an algorithm for theory assessment. *Third*, empirical factors are able to change the probability that such value judgments are correct (see SCRIVEN 1980; NAGEL 1986: pp. 143–153).

Fourth, to make empirical confirmability, instead of the ability to be subjected to the criticism of the scientific community, a necessary condition for objectivity would be to ignore the way that reasonable people behave. Reasonable people accumulate observations and inferences about judgments until the probability of those judgments is so great that they do not doubt them. They make assumptions when their inferences and evidence support them; they do not demand empirical confirmation for everything. Only if one were engaged in a search for certainty that transcends the possibility of error could one complain about well-supported scientific judgments that met the criteria for scientific objectivity just outlined. Since science has never claimed certainty that transcends the possibility of error, it is not reasonable to demand more than these criteria as standards for scientific objectivity (see SCRIVEN 1980: p. 286).

This new sense of “scientific objectivity” also seems plausible because it relies on the social and critical character of science, as POPPER (1950: pp. 403–406; 1962: p. 63; 1965: p. 56) realized, and as James, Wittgenstein and Wisdom suggested. If they are correct, then logical empiricists or universalists, as well as decision theorists, make too strict a demand on scientific objectivity by requiring confirmability.

3. Universality and scientific objectivity

The pluralists also make too strict demands on scientific objectivity by requiring that “objective” methodological rules or value judgments be infallible and universal. FEYERABEND (1977: p. 379; 1975: pp. 23, 28) maintains that “there is not a single rule, however plausible, and however firmly grounded in epistemology, that is not violated at some time or another”. Therefore, “there is only one principle that can be defended under all circumstances and in all stages of human development . . . the principle: ‘anything goes’.”

In searching for a certainty that appears to transcend the possibility of error, and in presupposing that objectivity requires infallibility and universality, FEYERABEND (1977: p. 368) appears to be presupposing that, since there is no perfect methodology, therefore all methodologies are equally bad: “Any procedure, however ridiculous, may lead to progress, any procedure however sound and rational, may get us struck in the mud.” Yet neither from the fact that all methods have been falsified historically nor from the fact that it is impossible that any method escape falsification, does it follow that all methods or methodological rules are equally bad (see KULKA 1977: pp. 279–280). Why not? *First*, the alleged historical falsifications of methods and the “irrationalist” advances in science provide only necessary, not sufficient, conditions for the correctness of claims that there is no universal rule of scientific method and that only a universal rule guarantees objectivity.

Second, as an attorney might point out, there is no obvious reason why rules must be exceptionless in order to be useful or rational. If rules are used for the purpose of scientific justification, then they ought not be exceptionless both because any type of justification is always complex and context-dependent (see HELLMAN 1979: p. 194; QUINE and ULLIAN 1970; HEMPEL 1965: p. 463), and because rules of scientific method need be merely acceptable to rational persons in the situation at hand, not universal (see SCRIVEN 1980: p. 277). Both scientific inference and legal inference establish that something is *prima facie* true, that it is reasonably probable, or that there is a presumption in its favor, not that it is infallibly true.

Third, great differences in scientific behavior are compatible with “objective” methodological rules and epistemic value judgments. Disagreements over rules do not mean either that there are no rules or that any rule is as good as another. Why not? Those who deny the existence of universal rules or values in science appear to do so because they fail to

distinguish three different questions: (1) Are there *general principles* (choose theories on the basis of explanatory power tested by prediction) that account for the rationality of science? (2) Are there *particular procedures*, or instantiations of the general principles, that account for the rationality of theory or paradigm choice? (3) Does science, *in fact*, always illustrate either the general principles or the specific procedures?

Feyerabend appears to assume that, if one answers questions (2) and (3) in the negative, then the answer to (1), the question before us, is also negative. This is false. Revolutionary debate, about question (2), does not jeopardize the rationality of science in the sense of suggesting there is no good answer to question (1). In fact, debate over question (2) must presuppose rationality in the sense of question (1), or the debate would be futile (see SIEGEL 1985: pp. 524–526; RUDNER 1966: pp. 4–5).

Another way to argue for this principle-versus-procedure or hierarchical conception of scientific rationality is to incorporate some insights from moral philosophy. In moral philosophy, as both natural-law philosophers and contemporary analysts such as R.M. Hare recognize, there is a hierarchy of methodological rules and value judgments, with different degrees of certainty appropriate to different levels of generality in the hierarchy. Science seems to have a similar hierarchy. In both science and in ethics, the most general rules are the most certain and the most universal, e.g., “choose the theory with the greater explanatory power”, or “do good and avoid evil”. The least general rules are the least certain and the least universal.

In both moral philosophy and in science, one must make a number of value judgments, especially at the lower levels of universality and generality, in order to interpret and to apply the rules from the most universal, most general level. In other words, in a specific situation, one must make very specific value judgments about what is “doing good”, and about what is “greater explanatory power”. Just because there is no algorithm, applicable to all situations, for deciding what is “doing good”, or what has “greater explanatory power”, does not mean that practice-based ethical or scientific rules and judgments are subjective or no better than any other possible judgment, since they are evaluated as *means* to the *end* or goal of explanatory power. Moreover, some methodological rules are better than others at reducing uncertainty, even though they do not guarantee infallibility. For example, it is clearly better, given the problem of diagnosing diabetes, to follow the methodological rule of performing a blood-sugar test than that of consulting a witchdoctor or examining sheep entrails (see TIBBETTS 1977: pp. 268–269).

Feyerabend misses both these points because he focuses on the *infallibility* of the scientific conclusions reached by means of methodological rules and epistemic value judgments, rather than on *prima facie* truth, or on what I have called “scientific objectivity”. This particular sense of objectivity relies on a number of insights of Popper, Wisdom, and Wittgenstein (see NEWELL 1986). It anchors objectivity with actions and practices rather than with an impossible, perhaps question-begging, notion of justification. In securing objectivity by means of the criticism of the scientific community, it presupposes that rationality and objectivity, in their final stages, require an appeal to particular cases as similar to other cases known to be correct, not an appeal to specific rules. This naturalistic appeal to cases, rather than to specific rules, is required (1) in order to avoid an infinite regress of justification, (2) because decisions about rules cannot rest on rules, and (3) because specific criteria would be too dogmatic to take account of counter instances.

More generally, if Feyerabendian arguments requiring that all objective methodological judgments be based on universal and stable rules were correct, and if Carnapian arguments requiring that all judgments in science be empirically confirmed were correct, and if these arguments were extended and used in other areas of epistemology, then they would invalidate most of our knowledge claims (see HELLMAN 1979: pp. 200–201).

What all these criticisms of the pluralists, the universalists, and the naturalists come down to is that they appear to have conceived of methodological rules and epistemic value judgments in science in a highly unrealistic way, either (respectively) as infallible and universal, as empirically confirmable, or they have confused general principles of method with specific procedures. I have argued that a more realistic way to conceive of such rules and judgments is in terms of “scientific objectivity” since, as even SCHEFFLER (1972: p. 369) recognizes, “objectivity requires simply the possibility of intelligible debate over the merits of rival paradigms”.

4. An example from hydrogeology

Although it would take extensive arguments and discussion of many historical examples to show that scientists ought to adopt both the notion of scientific objectivity I have proposed and explanatory power tested by prediction, as the guarantee of scientific rationality, let us look instead at

an example from hydrogeology, an example of what Hempel called "categorical judgments of value". It illustrates that, contrary to the universalists and pluralists, it is possible (in a particular context) to make objective, non-instrumental epistemic judgments of value, viz., to judge how to weight competing cognitive values.

Of course, there are a number of examples of categorical value judgments in physics. FRANKLIN's (1986: pp. 8–22, 35) analysis of the discovery of parity non-conservation, for instance, seems to provide an excellent example of how physicists affirmed parity conservation largely on grounds of simplicity. Beginning in 1956, however, because of initial experimental data on meson decays and because of its ability to solve the theta-tau puzzle, most of the physics community weighted the values of predictive power and internal coherence more highly than the value of simplicity and therefore affirmed parity non-conservation.

Rather than use a case from theoretical physics, however, I purposely have chosen an applied example, in which the scientific conclusions have consequences for human health and well-being. It graphically illustrates the point made by SCHEFFLER (1982: p. 5) and others (e.g., QUINE and ULLIAN 1970), that standards for scientific rationality and objectivity are important for controlling prejudice and ideology. If there is no scientific objectivity in the sense I have alleged, then science at the service of policy, as this example illustrates, will almost certainly be used for harm.

The example involves hydrogeological controversies that arose 25 years ago, in the course of scientists' attempts to ascertain subsurface migration rates for radionuclides at a proposed U.S. site for shallow land burial of radwaste. Geologists concluded that it would take 24,000 years for plutonium to migrate one-half inch at the site (USGS-P, n.d.; see WEISS and COLOMBO 1980: p. 5). Yet, only ten years after opening the facility, in 1973, plutonium and other radionuclides were discovered two miles offsite (MEYER 1975: p. 9). The geological predictions were wrong by *six orders of magnitude* and the site has become the world's "worst nuclear dump" (see NAEDELE 1979: pp. 1–3; BROWNING 1976: p. 43).

The geologists engaged in the site evaluation arrayed themselves in two opposed camps, industry/academia versus government. Scientists from several universities, EMCON Associates and NECO judged the site suitable for radwaste storage. Government scientists, from the U.S. Geological Survey (USGS) and the U.S. Environmental Protection Agency (EPA), judged the site unsuitable.

In 1974, ten years after the radwaste facility was opened, an EPA scientist (MEYER 1975) made a highly publicized claim that he had

discovered plutonium two miles offsite and that subsurface migration was responsible. Neither "side" in the controversy disputed the allegation that there was plutonium two miles offsite and that it had come from the site. The EMCON-NECO scientists, however, claimed that the plutonium had come from *surface* runoff from careless handling of radwaste spills; the USGS-EPA scientists held that the plutonium had gone offsite through *subsurface* migration from the burial trenches.

The surface contamination theory was reasonable both because external consistency supported it; because much of the plutonium discovered offsite was in streambed sediment near the surface (ZEHNER 1981: p. 58); and because the plutonium was associated with particulate matter (BLANCHARD *et al.* 1978: p. 29). Even if one hypothesized that plutonium was transported, deep beneath the surface, by means of some organic compound, all sides agreed that the agent responsible for the transport was wholly unknown (MOGHISSI 1976: p. 269; MEYER 1976: p. 2).

When the USGS and EPA scientists held to the subsurface migration theory, they admitted that their account contradicted scientific consensus that subsurface migration of plutonium was impossible (MEYER 1976: pp. 44–45). They said that this consensus was unable to explain an important anomaly, *viz.*, their finding plutonium deeper in the soil than it should have been (BLANCHARD *et al.* 1978: pp. 29, 1–5; ZEHNER 1981: pp. 104, 147; 1976: p. 256; see also KREY and HARDY 1970; McCLENDON 1975; PINDER *et al.* 1975). Hence they discounted the surface theory because of their adherence to the cognitive value of *internal coherence*. But how did they account for the alleged failure of the ion exchange that prevented plutonium migration? They did not discredit ion-exchange, but simply said that there was evidence, both that the mechanisms were unable to work as expected because of some intervening site variables (DEBUCHANNE 1976: p. 137; KENTUCKY SCIENCE AND TECHNOLOGY COMMISSION 1972: p. 8) and that studies alleging plutonium non-mobility were inapplicable to the site (MEYER 1976: pp. 44; ZEHNER 1976: p. 270; RHODES 1967; KNOLL 1969; NEUBOLD 1963; NEUBOLD *et al.* 1962; HALE and WALLACE 1970; WEISS and COLOMBO 1980: pp. xxii–xxiii, 121–135; see PRICE 1973; ROMNEY *et al.* 1970; MEYER 1976: pp. 45–46).

The divergent evaluations of plutonium non-migration theory apparently arose because different scientists attached different weights to cognitive values. The university/industry scientists weighted *external consistency* most heavily, while the government scientists weighted *internal coherence* most heavily. Yet, contrary to logical-empiricist or universalist claims, both groups of scientists appeared to be behaving rationally. If so, then

our example shows that, contrary to Hempel, Carnap and others, scientists as scientists make categorical judgments of value, and that the way they do so can determine the success or failure of their scientific conclusions. More importantly, the example shows that the situation itself suggests methodological rules that appear both rational and objective, in the sense that they are warranted by what we know about the situation and are able to be criticized and evaluated by the scientific community. One such rule might be to require studies of plutonium migration in fractured shale, like the site, since none had been done. Another rule might be to weight external consistency in proportion to the relevance of the evidence to the situation being examined. It is simply not true, as Feyerabend or Carnap might claim, that such rules, dictated largely by practice, are subjective. If not, then the scientific context itself, as the naturalists claim, provides clues for guaranteeing the scientific objectivity of the site-specific methodological rules and value judgments.

5. Conclusion

Where does this leave us? I argued that both the pluralists and the universalists fail because they presuppose an unrealistic ideal of scientific objectivity, one based on infallibility and universality. A more realistic notion of scientific objectivity, I argued, would lead to a position I call hierarchical naturalism, a stance midway between naturalism and universalism. Hierarchical naturalism recognizes that the universalists are correct in believing that there are at least general conceptual criteria for scientific rationality and objectivity, and that reason ought to alter scientific practice. It also recognizes that the naturalists are correct in believing that, for most useful methodological rules/epistemic value judgments, scientific rationality is largely a function of specific situations, and that scientific practice ought to alter reason.

Rather than take time to substantiate it myself, I cited Hempel's and McMullin's arguments for the claim that explanatory power, tested by prediction, provides a universal goal. Next I argued for a multi-leveled notion of scientific rationality and distinguished among principles or goals, procedures, and actual scientific practice.

My combination of naturalism and universalism, or what Laudan might call "reticulated" and "hierarchical" models of rationality, rests on the insight that scientific rationality and objectivity are more universal than the naturalists and pluralists claim and more complex than the universal-

ists appear to believe. This very complexity, however, gives hierarchical naturalism the ability to answer some of the main charges directed against typical variants of naturalism (e.g., the circle argument, the argument from norms, and the argument from relativism (see GIERE 1985)) and universalism (e.g., the plurality objection (see SIEGEL 1985: pp. 525–528)).

Against my specific position, hierarchical naturalism, there are at least seven main objections, all of which I believe can be answered. The first two have been formulated by Harvey Siegel, the third by Gary Gutting, the fourth and sixth by Phil Quinn. The fifth objection is suggested by one of Hempel's criticisms of Kuhn. The seventh objection has been stated most clearly by Vaughn McKim. These seven are: (1) To say that rules or goals of science are stable presupposes a realist view of science. (2) Since my account provides a normative view of scientific rationality and objectivity and is committed to a stable, universal goal of scientific inquiry, it cannot be called a *naturalistic* account. (3) Since I define "scientific objectivity" in terms of the criticism and debate of the scientific community, there appears to be no great difference between my view and that of Shapere and Feyerabend. (4) My account of scientific objectivity is "too thin". (5) My proposed goal of science, explanatory power tested by prediction, provides a trivial view of norms in science; as HEMPEL (1983: p. 87) suggests, desiderata are "imprecise constraints on scientific theory choice". (6) Alternatively, my proposed goal of science is too strong because it would require predictive power for many sciences not capable of providing it, e.g., anthropology. (7) Finally, it might be objected that my alleged example of a categorical value judgment in hydrogeology is really only an example of an instrumental value judgment.

Although space limitations prevent a full response to these seven objections, I shall briefly sketch the arguments that, if presented in full, would respond to them. *First*, use of the criterion of explanatory power, tested by prediction, does not commit me to realism since the entities having explanatory status may have only hypothetical or heuristic status. Moreover, pursuing an externalist position on scientific rationality, it is reasonable to argue *that* there is some universal, stable goal of science, without arguing *why* this stability is the case (NAGEL 1986: p. 81; GOLDMAN 1980: pp. 27–51; CHISHOLM 1982: pp. 61 ff.).

Second, I am not a universalist in that I am not committed to purely *a priori* rules of scientific method and because I believe that my goal (explanatory power as tested by prediction) underdetermines all specific

methodological rules; hence the specific rules need to be dictated largely by the particular situation. *Third*, because the specific rules need to be dictated largely by the particular situation does not mean that my account of scientific objectivity is no different from that of Shapere and Feyerabend. Contrary to them, I claim that there is a universal, stable goal of science, explanatory power as tested by prediction, and that scientific rationality, as defended in the essay, is a necessary condition for scientific objectivity.

Fourth, my notion of scientific objectivity is not too thin because (A) it presupposes a notion of scientific rationality dependent upon explanatory and predictive power; (B) even Phil Quinn suspects that the goal of explanatory power, tested by prediction, provides *too strong* a requirement for scientific rationality (see objection six); (C) any stronger definition of scientific objectivity seems likely to fail either because it might beg the realism question, or because it might presuppose knowledge we do not have. Because every situation in science is different, it is virtually impossible to specify, ahead of time, what an objective representation of some situation might be.

Fifth, the universal goal, explanatory power as tested by prediction, does not provide a trivial view of norms in science, both because it rejects most common versions of naturalism and because it provides an answer to question (1) mentioned earlier in the text, "Are there *general principles* that account for the rationality of science?" *Sixth*, another reason for believing that explanatory power, as tested by prediction, is not a trivial goal of science is that at least some philosophers of science believe that prediction is too strong a goal for many sciences, such as anthropology. However, (A) anthropologists need not attain predictive power, but only admit that it is a characteristic goal of their activities; (B) without this goal, one could not test a scientific explanation and adequately secure the empirical foundations of science; (C) if the goal of predictive accuracy were irrelevant to a particular science, then it would be likely that the alleged science was not science, but something related to it, e.g., natural history. Finally, (D) I have not argued that predictive power is an appropriate goal of science; as explained earlier, I have left this task to McMullin and Hempel.

Seventh, my example of a categorical value judgment in science, viz., weighting internal coherence more heavily than external consistency in evaluating theories of plutonium migration, is not really only an example of an instrumental value judgment. Even if the scientists aimed to site/to reject the radwaste facility, and they simply chose whatever epistemic

goal, e.g., external consistency, was instrumental to their pragmatic end, at least three important points ought to be noted. (1) In the examples discussed in this essay, the scientists' own statements provide strong evidence for their use of categorical value judgments about internal and external consistency. (2) There is no strong evidence that the ultimate goals of the two groups of scientists were, respectively, to site and not to site the facility; admittedly, however, their pursuing such goals is highly plausible, given their vested interests. (3) Even if it could be established that the two groups of scientists pursued such political/economic goals, and that the values discussed in my examples were purely instrumental to these goals, there are numerous other examples of categorical value judgments in science, e.g., regarding parity non-conservation and regarding the postulation of the neutrino earlier in this century. Indeed, once the theoretical controversy over plutonium migration is separated from its application to a radwaste siting dispute, it could be used to establish this same point. Unfortunately, there is no space here to spell out either alternative examples or the arguments supporting responses to this and the other objections. These argument sketches should be enough to suggest, however, that my notions of scientific rationality and objectivity deserve further investigation, even though the account presented here leaves much of the epistemological work still to be done.

Acknowledgement

The author is indebted to Gary Gutting, Vaughn McKim, Ernan McMullin, Harvey Siegel and Phil Quinn for comments on an earlier version of this essay. Whatever errors that remain are the sole responsibility of the author.

References

- BLANCHARD, R., MONTGOMERY, D., KOLDE, H. and GELS, G., 1978, *Supplementary Radiological Measurements at the Maxey Flats Radioactive Waste Burial Site — 1976–1977*, EPA-520/5-78-011, Montgomery, Alabama, Office of Radiation Programs, U.S. Environmental Protection Agency.
- BROWNING, F., 1976, *The Nuclear Wasteland*, *New Times* 7, pp. 43–47.
- CARNAP, R., 1950, *Logical Foundations of Probability* (University of California Press, Chicago).

- CARNAP, R., 1952, *The Continuum of Inductive Methods* (University of Chicago Press, Chicago).
- CARNAP, R., 1967, *The Logical Structure of the World and Pseudoproblems in Philosophy* (University of California Press, Berkeley).
- CAWS, P., 1967, *Science and the Theory of Value* (Random House, New York).
- CHISHOLM, R., 1982, *The Foundations of Knowing* (University of Minnesota Press, Minneapolis).
- DAVIDSON, D., 1984a, *Inquiries into Truth and Interpretation* (Clarendon Press, Oxford).
- DAVIDSON, D., 1984b, *A Coherence Theory of Truth and Knowledge*, in: Henrich, ed., *Kant oder Hegel* (Stuttgart).
- DEBUCHANANNE, G.D., 1976, *Statement*, U.S. Congress.
- FEYERABEND, P. 1975, *Against Method* (New Left Bookstore, London).
- FEYERABEND, P., 1977a, *Marxist Fairytales from Australia*, *Inquiry* 20, pp. 372–397.
- FEYERABEND, P., 1977b, *Changing Patterns of Reconstruction*, *Br. J. Philos. Sci.* 28, pp. 351–369.
- FEYERABEND, P., 1978, *Life at the LSE*, *Erkenntnis* 13, pp. 297–304.
- FEYERABEND, P., 1980, *How to Defend Society Against Science*, in: Klemke, Hollinger and Kline, eds., *Introductory Readings in Philosophy of Science* (Prometheus, Buffalo).
- FRANKLIN, A., 1986, *The Neglect of Experiment* (Cambridge University Press, New York).
- GIERE, R., 1985, *Philosophy of Science Naturalized*, *Philos. Sci.* 52, pp. 331–357.
- GOLDMAN, A., 1980, *The Internalist Conception of Justification*, in: P.A. French, ed., *Midwest Studies in Philosophy*, vol. 5, *Studies in Epistemology* (University of Minnesota Press, Minneapolis).
- HALE, V. and WALLACE, A., 1970, *Effects of Chelates on Uptake of Some Heavy Metal Radionuclides from Soil by Bush Beans*, *Soil Sci.* 109, p. 26 ff.
- HELLMAN, G., 1979, *Against Bad Method*, *Metaphilosophy* 10, pp. 190–202.
- HEMPEL, C., 1965, *Aspects of Scientific Explanation* (Free Press, New York).
- HEMPEL, C., 1979, *Scientific Rationality*, in: T. Geraets, ed., *Rationality Today* (University of Ottawa Press, Ottawa).
- HEMPEL, C., 1980, *Science and Human Values*, in: Klemke, Hollinger and Kline, eds. *Introductory Readings in Philosophy of Science* (Prometheus, Buffalo).
- Hempel, C., 1983, *Valuation and Objectivity in Science*, in: Cohen and Laudan, eds., *Physics, Philosophy, and Psychoanalysis* (Reidel, Dordrecht).
- KENTUCKY SCIENCE AND TECHNOLOGY COMMISSION, 1972, *Technical Review of the Maxey Flats Radioactive Waste Burial Site*, unpublished report, Kentucky Department of Human Resources, Frankfort, Kentucky.
- KLEMKE, E., HOLLINGER, R. and KLINE, A., eds., 1980, *Introductory Readings in the Philosophy of Science* (Prometheus, Buffalo).
- KNOLL, K., 1969, *Reactions of Organic Wastes and Soils*, U.S. Atomic Energy Commission Report BNWL-860.
- KREY, P. and HARDY, E., 1970, *Plutonium in Soil Around the Rocky Flats Plant*, U.S. Atomic Energy Commission Report HASL-235.
- KULKA, T., 1977, *How Far Does Anything Go? Comments on Feyerabend's Epistemological Anarchism*, *Philos. Social Sci.* 7, pp. 277–287.
- LAUDAN, L., 1984, *Science and Values* (University of California Press, Berkeley).
- MARTIN, M., 1972, *Concepts of Science Education* (Scott, Foresman, Glenview, IL).
- McCLENDON, H.R., 1975, *Soil Monitoring for Plutonium at the Savannah River Plant*, *Health Phys.* 28, 347 ff.
- McMULLIN, E., 1983, *Values in Science*, in: P. Asquith, ed., *PSA 1982*, vol. 2 (Philosophy of Science Association, East Lansing).
- McMULLIN, E., 1986, *The Shaping of Scientific Rationality: Construction and Constraint*, Inaugural Address in the O'Hara Chair of Philosophy at the University of Notre Dame.

- MEYER, G., 1975, *Maxey Flats Radioactive Waste Burial Site: Status Report*, unpublished report, Advance Science and Technology Branch, U.S. Environmental Protection Agency.
- MEYER, G., 1976, *Preliminary Data on the Occurrence of Transuranium Nuclides in the Environment at the Radioactive Waste Burial Site, Maxey Flats, Kentucky*, U.S. Environmental Protection Agency, Office of Radiation Programs, Washington, D.C.
- MOGHISSI, A., 1976, *Letter to H. Holton*, NECO, February 17, 1976, in: U.S. Congress, pp. 269-270.
- NAEDELE, W.F., 1979, *Nuclear Grave is Haunting Kentucky*, Philadelphia Bulletin, May 17, 1979, in: U.S.G.S., Maxey Flats—Publicity, vertical file, Louisville, Kentucky Water Resources Division, U.S. Division of the Interior.
- NAGEL, E., 1961, *The Structure of Science* (Harcourt Brace, New York)
- NAGEL, T., 1986, *The View From Nowhere* (Oxford University Press, New York).
- NEUBOLD, P., 1963, *Absorption of Plutonium-239 by Plants*, GB Agr. Res. Council, ARRCL-10.
- NEUBOLD, P. and MERCER, E., 1962, *Absorption of Plutonium-239 by Plants*, GB Agr. Res. Council, ARRCL-8.
- NEWELL, R., 1986, *Objectivity, Empiricism, and Truth* (Routledge and Kegan Paul, New York).
- PINDER, J.E., et al., 1975, *A Field Study of Certain Plutonium Contents of Old Field Vegetation and Soil Under Humid Climatic Conditions*, Proceedings, Radiology Symposium, Corvallis, Oregon.
- POPPER, K., 1950, *The Open Society and its Enemies* (Princeton University Press, Princeton, NJ).
- POPPER, K., 1962, *Conjectures and Refutations* (Basic Books, New York).
- POPPER, K., 1965, *The Logic of Scientific Discovery* (Harper, New York).
- PRICE, K.R., 1973, *A Review of Transuranic Elements in Soils, Plants, and Animals*, J. Environ. Qual. 2, p. 62 ff.
- PUTNAM, H., 1981, *Reason, Truth, and History* (Cambridge University Press, New York).
- QUINE, W.V. and ULLIAN, J.S., 1970, *The Web of Belief* (Random House, New York).
- RHODES, D., 1967, *Absorption of Plutonium by Soil*, Soil Sci. 84, p. 465 ff.
- ROMNEY, E., et al., 1970, *Persistence of Plutonium in Soil, Plants, and Small Mammals*, Health Phys. 22, p. 487 ff.
- RORTY, R., 1986a, *Pragmatism, Davidson, and Truth*, in: E. LePore, ed., *Essays on "Inquiries into Truth and Interpretation"* (Blackwell, Oxford).
- RORTY, R., 1986b, *Back to the Demarcation Problem*, a paper presented at the University of Notre Dame, at a conference titled "The Shaping of Scientific Rationality", unpublished manuscript.
- RUDNER, R., 1966, *Philosophy of Social Science* (Prentice-Hall, Englewood Cliffs, NJ).
- RUDNER, R., 1980, *The Scientist qua Scientist Makes Value Judgments*, in: Klemke, Hollinger and Kline, eds., *Introductory Readings in Philosophy of Science* (Prometheus, Buffalo).
- SCHEFFLER, I., 1972, *Discussion: Vision and Revolution: a Postscript on Kuhn*, Philos. Sci. 39, pp. 366-374.
- SCHEFFLER, I., 1982, *Science and Subjectivity* (Hackett, Indianapolis).
- SCHLICK, M., 1959, *The Foundations of Knowledge*, in: A.J. Ayer, *Logical Positivism* (Free Press, New York).
- SCRIVEN, M., 1974, *The Exact Role of Value Judgments in Science*, in: R. Cohen and K. Schaffner, *Proceedings of the 1972 Biennial Meeting of the Philosophy of Science Association* (Reidel, Dordrecht).
- SCRIVEN, M., 1980, *The Exact Role of Value Judgments in Science*, in: Klemke, Hollinger and Kline, eds., *Introductory Readings in Philosophy of Science* (Prometheus, Buffalo).

- SELLARS, W., 1967, *Philosophical Perspectives* (Charles Thomas, Springfield, Illinois).
- SHAPERE, D., 1984, *Reason and the Search for Knowledge* (Reidel, Dordrecht).
- SIEGEL, H., 1985, *What is the Question Concerning the Rationality of Science?* *Philos. Sci.* 52, pp. 517–537.
- TIBBETTS, P., 1977, *Feyerabend's 'Against Method': The Case for Methodological Pluralism*, *Philos. Social Sci.* 7, pp. 268–269.
- U.S. CONGRESS, 1976, *Low-Level Radioactive Waste Disposal*, Hearings before a subcommittee of the Committee on Government Operations, House of Representatives, 94th Congress, Second Session, February 23, March 12, and April 6, 1976 (U.S. Government Printing Office, Washington, DC).
- U.S. GEOLOGICAL SURVEY (n.d.) *Maxey Flats: Publicity*, vertical file, Water Resources Division, U.S. Division of the Interior, Louisville, Kentucky.
- WEISS, A. and COLUMBO, P., 1980, *Evaluation of Isotope Migration—Land Burial*, NUREG/CR-1289 BNL-NUREG-51143, U.S. Nuclear Regulatory Commission, Washington, DC.
- ZEHNER, H., 1976, *Notes in the Margin of U.S. Congress*, 1976, Water Resources Division, U.S. Division of the Interior, Louisville, Kentucky.
- ZEHNER, H., 1981, *Hydrogeologic Investigation of the Maxey Flats Radioactive Waste Burial Site*, Fleming County, Kentucky, Open-File Report, draft, USGS, Louisville, Kentucky.

This Page Intentionally Left Blank

7

**Foundations of Probability
and Statistical Inference**

This Page Intentionally Left Blank

THE INTERFACE BETWEEN STATISTICS AND THE PHILOSOPHY OF SCIENCE

I.J. GOOD

*Dept. of Statistics, Virginia, Polytechnic Institute and State University,
Blacksburg, VA 24061, USA*

My topic is the interface between statistics and the philosophy of science; that is, the influences that each has had or might have on the other. Many people have contributed to this topic but I shall largely review the writings of I.J. Good because I have read them all carefully. These influences are related to the semi-quantitative ideas that emerge from an informal Bayesian approach, jestingly called Doogian.

A longer version of this paper will be published, with discussions, in *Statistical Science*.

Among the topics that I shall touch upon are probability, surprise, rationality, corroboration or weight of evidence, explanation, induction, probabilistic causality, and a Bayes/non-Bayes compromise.

My discussions belong to a field that can be called the mathematics of philosophy or probabilistic philosophy. The approach is often only semi-quantitative because of the difficulty or impossibility of assigning precise numbers to the probabilities. Some people will argue that it is misleading to use precise-looking formulae for concepts that are not precise, but I think it is more leading than misleading because a formula encapsulates many words and provides a goal that one can strive towards by sharpening one's judgments. Also it is easier to make applications to statistics if one has a formula. A semi-quantitative theory should be consistent with a good qualitative theory. For example, I think this applies basically to my theory of probabilistic causality (GOOD 1961/1962, 1984/1985, 1987a) in relation to the more qualitative theory of SUPPES (1970). A reader who holds in mind the present paragraph will not be misled by the apparent precision of the formulae.

Probability

POISSON (1837: p. 2) made a clear distinction between two kinds of probability which may be called epistemic and physical (see GOOD 1986a, for further discussion). Epistemic probability can be either subjective (= personal) or logical and finer classifications have been given (KEMBLE 1942; GOOD 1959, 1966; FINE 1973).

Poisson assumes that the probability of an event is different for different people only because they have different information. This seems to imply that $P(A | B)$ is the same for everybody so Poisson must have had credibility (= logical probability) in mind. Subjective probability was regarded as the most basic kind by RAMSEY (1926/1964) and DE FINETTI (1937/1964), and in books by GOOD (1950) and by SAVAGE (1954). Early modern books on credibility were written by KEYNES (1921), JEFFREYS (1939) and CARNAP (1950), though all three of these authors later became more sympathetic to the use of subjective probability than they were when they wrote those books.

I doubt whether credibility can ever be given a convincing precise numerical meaning unless the information has symmetry properties, or if the sample is very large, but I believe it is a useful fiction to assume that credibility has sharp values even when there is no sample, and I think it is mentally healthy for you to think of your subjective probabilities as estimates of credibilities. (The concept of useful fictions was developed by Jeremy Bentham in the early nineteenth century: see OGDEN (1959).) Physical probability too is a useful fiction even if the world is deterministic, just as pseudorandom numbers are regularly used by statisticians as if they were strictly random. (See both indexes of GOOD (1983e) under "determinism".) De Finetti proved a theorem that can be interpreted as saying that a person who has sharp subjective probabilities that are consistent with the axioms behaves *as if* physical probabilities exist (although de Finetti believed they do not exist) and these physical probabilities have unique subjective probability distributions. The theorem can also be interpreted as saying that solipsism cannot be strictly disproved. De Finetti did not express the theorem in either of these ways. It is an excellent example of a theorem at the interface between philosophy and statistics. For a simple exposition of de Finetti's theorem see GOOD (1965: pp. 12–14, 22–23). For its relationship to the non-disprovability of solipsism see GOOD (1983e: pp. 93, 154), where further references are mentioned.

KEYNES (1921) argued that credibilities should be regarded as partially ordered. GOOD (1950) adopted the same view for subjective probabilities, but sometimes it is a good enough approximation to think of a probability as having a sharp numerical value. I think the simplest satisfactory theory of partially-ordered subjective probability, or any other well-founded scientific theory, is one based on axioms, rules of application and *suggestions*. I listed 27 suggestions in GOOD (1970/1971: pp. 124–127) and called them the Priggish Principles.

The use of partially-ordered probabilities can be regarded as a kind of “formalization of vagueness”. It differs from the theory of fuzzy sets which deals with “degrees of belonging” to a set or, as it might be expressed, with “degrees of meaning” (GOOD 1950: p. 1). For example, it is more meaningful to say that a man has a beard if he resembles a religious leader, such as Christ, Santa Claus or Karl Marx, than if his chin is merely fuzzy.

When all your prior probabilities are sharp you are a strict Bayesian, whereas, when all upper and lower prior probabilities are 1 and 0 respectively, you are a strict non-Bayesian. Because I believe that subjective probabilities are only partially ordered I am forced into a Doogian intermediate position. I am forced to look for compromises between Bayesian and non-Bayesian methods and especially ways in which a somewhat Bayesian outlook can shed light on and improve so-called non-Bayesian methods.

I regard it as acceptable to use seemingly non-Bayesian methods *except when they are seen to contradict your own judgments of probabilities etc. in a given application*, the axioms of subjective probability being assumed. Whether you arrive at a contradiction will depend partly on how much thought you give to the matter. The type II principle of rationality recommends that you should allow for the cost of thinking and calculation when trying to apply the type I principle, namely the maximization of expected utility. Thinking will often cause you to change your mind; that is why dynamic probabilities are relevant: see, for example, GOOD (1977a).

You can make probability judgments about the accuracy of your own judgments, and this leads to a hierarchical Bayesian approach in statistics, not necessarily restricted to only two levels. This approach is at least an aid to the judgment. It was exemplified by a so-called type II minimax procedure in GOOD (1952). Later it led to an adequate Bayesian significance test for multinomials (GOOD 1965, 1967; GOOD and CROOK 1974;

LEONARD 1977). The basic idea is to use prior distributions that contain parameters known as hyperparameters, and these can be assigned hyperpriors (see GOOD 1979/1981).

My guess about the future of statistics is that it will be a compromise between hierarchical Bayesian methods and methods that seem superficially to be non-Bayesian.

Induction

By scientific induction I mean changing the probability of hypotheses in the light of evidence or observations and thereby also changing the probabilities of future observations. The problem is partly solved by means of Bayes's theorem. Some people call the *formulation* of hypotheses "induction", but I prefer the obvious name *hypothesis formulation* for that activity. Sometimes hypotheses can be formulated automatically by maximizing entropy (see GOOD 1963).

The estimation of physical probabilities of multinomial (or binomial) categories is of course a contribution to the problem of scientific induction. In particular the hierarchical Bayesian method was used explicitly for this purpose by GOOD (1983a,b). A qualitative consequence of the hierarchical approach, and of the calculations, was that "induction to the next trial" is much more reliable than "universal induction" or "induction to all future trials", and I think most people would agree with this conclusion without detailed analysis.

The first quantitative contribution to scientific induction was Laplace's Law of Succession. For example, if you have seen n swans in England and they have all been white, and if you assume no other knowledge, then the odds are $n + 1$ to 1 that the next one chosen at random will be white according to Laplace's law, or $2n + 1$ to 1 on the basis of an "invariant" prior for the binomial parameter that was proposed by H. Jeffreys and independently by W. Perks (their priors differed for multinomials). If the conditions change in a substantial manner, for example, if the next observation is made in Australia, you cannot be so sure, and in fact there are *black* swans there. This kind of thinking, by those in charge, would have prevented the Challenger disaster.

A special case of a hypothesis is that a specific word has a specific meaning or class of meanings, and this hypothesis is made more probable if you look the word up in a dictionary and also observe how the word is used. This applies to every word in the language including "induction"

itself so if someone tells me he does not believe at all in probabilistic induction, as I understand the expression, for all I know he is asserting that the moon is made of gorgonzola or that pigs eat purple people (BLACK 1967; GOOD 1981b). It is like a non-dreaming solipsist trying to convince other people he is right.

Induction is closely related to the concept of weight of evidence and I discuss this concept next.

Weight of evidence

Let H denote a hypothesis, such as that an accused person is guilty, and let E denote some evidence such as that presented by a specific witness. We ask how should we define $W(H : E | G)$, the weight of evidence in favour of H provided by E when background knowledge G is regarded as given or previously taken into account. It is natural to assume that the new evidence converts the prior probability into its posterior probability, that is, that $P(H | E \& G)$ is a mathematical function of $P(H | G)$ and of the weight of evidence. Moreover $W(H : E | G)$ should depend only on (i) the probability of E given that the accused is guilty, and (ii) the probability of E given that he is innocent, that is on $P(E | H \& G)$ and $P(E | \bar{H} \& G)$ where the bar denotes negation. These desiderata lead to the conclusion that $W(H : E | G)$ must be a monotonic function of the Bayes factor $P(E | H \& G) / P(E | \bar{H} \& G)$ and we may well take the logarithm of the Bayes factor as our explicatum because this leads to desirable additive properties of the kind assumed by the goddess Themis (compare GOOD 1968b, 1984b). In fact

$$W[H : (E \& F)] = W(H : E) + W(H : F | E). \quad (1)$$

I have taken G for granted to simplify the appearance of the formula. When E and F are independent given H and also given \bar{H} , this formula reduces to

$$W(H : E \& F) = W(H : E) + W(H : F).$$

It was pointed out by WRINCH and JEFFREYS (1921), in a slightly different notation, that

$$\frac{P(E | H \& G)}{P(E | \bar{H} \& G)} = \frac{O(H | E \& G)}{O(H | G)} \quad (2)$$

the ratio of the final (posterior) to the initial (prior) odds. Thus W is the additive change in the log-odds of H by virtue of E .

It is best to think of the Bayes factor as defined by the right-hand side of (2), that is, as the factor by which the initial odds of H are multiplied to obtain the final odds. It is convenient that this factor is equal to the left side because this can be evaluated independently of the initial probability of H which can be especially difficult to judge. I conjecture that most juries are able to judge *final* probabilities of guilt better than *initial* probabilities because in ordinary affairs final probabilities are more important than initial ones so we think about them more.

Because the left-hand side of (2) sometimes reduces to a simple likelihood ratio we can regard a Bayes factor as part of the interface between Bayesian and less philosophical non-Bayesian statistics. Ordinary (non-Bayesian) likelihood is also part of this interface.

The technical concept of weight of evidence, because it captures the intuitive concept so well, should be of interest in legal matters (GOOD 1986c), and is already of interest for medical diagnosis, especially differential diagnosis (between two diseases) (see, e.g., GOOD and CARD 1971; CARD and GOOD 1974; SPIEGELHALTER and KNILL-JONES 1984).

The concept of a *unit* of weight of evidence is due to TURING (1941). He talked of bans, decibans and natural bans, the latter when natural logarithms are used. The deciban resembles the decibel in acoustics, being about the smallest weight of evidence perceptible to the human mind. Turing's name for a weight of evidence was "score" or "decibannage".

PEIRCE (1878) almost anticipated the best formal concept of weight of evidence but his definition applies only if the initial odds of H are 1 or "evens", that is, if $P(H | G) = \frac{1}{2}$. In this special case the weight of evidence is equal to the posterior log-odds. JEFFREYS (1939) also nearly always assumes that $O(H) = 1$ in spite of his earlier work. This was because in his book he was trying to be a credibilist, especially in the first edition. POISSON (1837: Chap. V) also came close to the formal concept (see GOOD 1986a: p. 167).

Weight of evidence can be regarded as a quasi-utility or epistemic utility, that is, as a substitute for utility when the actual utilities are difficult to estimate. (A quasi-utility can be defined as an additive epistemic utility.) Just as for money, diminishing returns eventually set in; for example, in a court of law, if the weight of evidence in favour of guilt or innocence becomes overwhelming there is little point in seeking further evidence, especially if it is expensive. The same principle applies

in scientific or medical research or even in a game of chess (where evolving or dynamic probabilities are relevant (see GOOD 1968b; and especially 1977a). But the effect of diminishing returns can often be ignored. When this is done we naturally bring in the concept of *expected* weight of evidence, which, in discriminating between two multinomials, leads to an expression of the form

$$\sum_i p_i \log (p_i/q_i). \quad (3)$$

This, or its general form (continuous or mixed), is often called cross-entropy. Such expressions were used by GIBBS (1875/1906/1961: p. 163), somewhat implicitly, in statistical mechanics, and in statistics by a number of later authors. For many references see GOOD (1983/1985), CHRISTENSEN (1983: Chap. 1), and the indexes of GOOD (1983e) under "Weight of Evidence". Ordinary entropy is effectively minus a special case, namely when q_i has the same value for all i . In the design of an experiment for estimating a parameter it might be reasonable to maximize the expected cross-entropy; but to minimize the cross-entropy when doing the estimation after the experiment is done (cf. GOOD 1968a). This is because, according to a theorem due to WALD (1950: p. 18), a minimax solution is a Bayes solution that uses the *least* favourable prior. Minimax solutions are not optimal but they have the merit of invariance under changes of variables (see also GOOD 1955/1956, 1969; LINDLEY 1956).

Tail probabilities or *P*-values

In statistical practice a small *P*-value such as $\frac{1}{50}$ is usually regarded as evidence against the "null hypothesis" H , and there is a temptation to think that any fixed value, say $P = 0.031$ (which is not the same assertion as that $P < 0.05$ (see GOOD 1950: p. 94n)) conveys the same amount of evidence against H on all occasions, at any rate if we are careful to use either single tails or double tails, depending on circumstances. This temptation must be resisted for several different totally convincing reasons. Some of these reasons are mentioned in my paper on hypothesis testing (GOOD 1981a), and I shall not repeat them here. Elsewhere I have given a brief discussion of what it means to say that a theory is *true*. Here I would like to mention that a very simple argument can be given, without mentioning Bayes or Neyman and Pearson, to prove conclusively the diminishing significance of a fixed *P*-value when a sample size is increased (GOOD 1983c). Indeed, *given a fixed statistical model*, a fixed

P -value, however small, can *support* the null hypothesis if the sample size is large enough and if the mathematical model is *sufficiently reliable*.

In several situations the Bayes factor against a sharp null hypothesis is roughly proportional to $1/(P\sqrt{N})$ (see JEFFREYS 1939: Appendix 1; GOOD 1983e: p. 143). One way to understand this is that the prior measures of reasonable sets of non-null hypotheses, such as $97\frac{1}{2}\%$ confidence intervals, shrink roughly proportionally to $1/\sqrt{N}$. I have accordingly suggested (GOOD 1982, 1984a, 1984c, 1984d) that P -values, if you *must* use them, should be standardized to a fixed sample size, say $N = 100$, by replacing P by

$$\min\left(\frac{1}{2}, P\sqrt{N/100}\right) \quad (4)$$

(when $N > 10$) and calling it a P -value *standardized to sample size 100*. Standardized P -values exemplify the concept of a Bayes/non-Bayes compromise. Several other examples, and historical comments, can be found, for example, in a recent encyclopedia article on scientific method and statistics (GOOD 1985/1987). Note that when you are sure that there are only two “simple statistical hypotheses”, there is little point in using P -values rather than Bayes factors.

The combination of P -values in parallel

Let P_1, P_2, \dots be some P -values obtained by distinct tests, but *based on the same data*. I call these “tests in parallel”. A dishonest experimenter might choose the smallest or largest of these depending on whether he is bribed or intimidated to disprove or to support the null hypothesis. A rule of thumb that seems to appeal even to non-Bayesians is to replace these P -values by their harmonic mean or perhaps by a weighted harmonic mean. This proposal has an informal Bayesian justification, and is a nice example of a Bayes/non-Bayes compromise (GOOD 1958).

The choice of a criterion for a significance test

An early example of a Bayes/non-Bayes compromise, understood explicitly as such, was related to the choice of a criterion for a significance test (GOOD 1957: p. 863). The proposal was to compute a Bayes factor (or equivalently a weight of evidence), based on a Bayesian model in which

you do not necessarily have much confidence, and then to treat this Bayes factor merely as a criterion in the “Fisherian” manner so to speak by obtaining its distribution given the null hypothesis. (I sometimes refer to the use of P -values as Fisherian to distinguish this usage from the acceptance-rejection procedure of Neyman and Pearson and from strict Bayesian methods. But P -values have a history dating back for two centuries. The idea of finding the distribution of a Bayes factor, given the null hypothesis, is now practicable up to a point by simulation (Good 1986d).) Fisher used to select criteria for significance tests without any explicit formal principle, but based on “common sense”, although he had explicit principles for estimation problems. His common sense was undoubtedly based on some vague non-null composite hypotheses, in fact he said (FISHER 1955: p. 73), in relation to P -values, that “The deviation [might be] in the direction expected for certain influences *which seemed to me not improbable . . .*” (my italics). Note the personal Bayesian tone here and he also refers to “the tester’s state of mind”. I wonder where this explicit subjectivism first occurred in Fisher’s writings. Did it occur before the revival of the modern subjectivistic movement? Strict Bayesians and Neyman–Pearsonians have to select precise non-null hypotheses, although in reality there is nearly always some vagueness in the real world. How much should be formalized and how much should be left vague depends partly on personal judgment.

Note that the Neyman–Pearson–Wilks “likelihood ratio”, a ratio of maximum likelihoods, can be regarded as a crude approximation to a Bayes factor for a *very bad* Bayesian model, yet it works well as a significance criterion. The basic idea is that, if you cannot evaluate an integral, work instead with the maximum of the integrand without even allowing for the curvature of the integrand at its maximum! This crude idea also leads to the use of maximum likelihood as another example of a Bayes/non-Bayes compromise.

When the number of parameters is large this informal Bayesian justification of maximum likelihood estimation is liable to break down, and then, I believe, the method of maximum likelihood becomes unacceptable. A very good example is the estimation of a probability density function f given a finite sample of observations x_1, x_2, \dots, x_N . In this case the number of *parameters* is infinite and the maximum likelihood estimate consists merely of one N th of a Dirac function at each observation. This disaster can be avoided by using the method of maximum penalized likelihood in which the log-likelihood $\sum_i \log f(x_i)$ is penalized by subtracting from it a *roughness penalty* $\Phi(f)$ such as $\beta \int [(\sqrt{f})'']^2 dx$

where β is called a hyperparameter or smoothing parameter (GOOD and GASKINS 1971, 1972, 1980; GOOD and DEATON 1981; LEONARD 1978). The method of maximum penalized likelihood was described by GOOD and GASKINS (1972) as a wedding between Bayesian and non-Bayesian methods because one can either regard $\exp(-\Phi)$ as proportional to a prior density (possibly improper) in function space or else the whole procedure can be regarded as a common-sense *ad hoc* non-Bayesian adjustment of maximum likelihood estimation to save it from disaster. (A special feature of the Bayesian interpretation is that it leads to a way of evaluating bumps.) A similar penalizing of a log-likelihood was also suggested earlier by GOOD (1963: p. 931), the idea being to maximize a linear combination of log-likelihood and entropy. When there is no sample this suggestion reduces to the method of maximum entropy, so the proposal was a generalization of that method.

For all these procedures it is necessary to choose the hyperparameter, procedural parameter, or smoothing parameter. Methods were given by Good and Gaskins but their reliability needs to be investigated by further simulation methods. One could also assume a hyperprior for the hyperparameter. If a sample is large then the smoothing parameter can be reliably estimated by using the old-fashioned split-sample method or by means of the modern modifications called cross-validation or predictive sample reuse, though these methods can be expensive.

The theory of significance tests, based on P -values, cannot be entirely separated from the theory of estimation of parameters. Thus FISHER (1955) said “. . . in the theory of estimation we consider a continuum of hypotheses each eligible as a null hypothesis, and it is the aggregate of frequencies calculated from each possibility in turn as true — including frequencies of error [P -values], therefore only of the ‘first kind’, without any assumptions of knowledge *a priori* — which supply the likelihood function, fiducial limits, and other indications of the amount of information available.” In this way Fisher was able to subsume the concept of errors of the second kind under those of the first kind. This P -value function is a continuous form of all possible confidence intervals, although Fisher might have deliberately avoided this mode of expression! It is not surprising that PEARSON (1955) said, in response, that “. . . I do not think that our position in some respects was or is so very different from that which Professor Fisher himself has *now* reached” (my italics). Another implication of Fisher’s remark is in suggesting the notion of a continuum of hypotheses possibly forming an “onion”, or part of an onion, surrounding the null hypothesis, these hypotheses being “more

non-null” when further out, and the inner core being virtually the null hypothesis.

Fisher mentioned fiducial limits in the quoted passage, so I will remind you *en passant* that Fisher’s fiducial argument is fallacious and the reason he made a mistake was simply because he did not use a notation for conditional probability. FISHER (1955: p. 74) says “He [Neyman] seems to claim that the statement (a) ‘ θ has a probability of 5 per cent of exceeding T ’ is a different statement from (b) ‘ T has a probability of 5 per cent of falling short of θ ’.” In my opinion the error was Fisher’s because only one of the statements should be made conditional on T . Bad notations and terminology tempt people into making substantial errors. An example was Carnap’s use of “confirmation” for logical probability, a usage that still causes confusion among philosophers of science. The ordinary English meaning of confirmation is much closer to weight of evidence than to probability. I predict that the misuse of the term “confirmation” will continue until the year 2002.

JEFFREYS (1939) showed that in some circumstances the use of the fiducial argument was equivalent to assuming a specific Bayesian prior, usually “improper”, that is, integrating to infinity instead of 1. According to STIGLER (1986: pp. 91, 102–104) the fiducial argument was foreshadowed by Thomas Simpson in 1755 and its relation to inverse probability was recognized, but only implicitly, by Laplace. The error in the exposition of the fiducial argument by Fisher (1955) *together with the psychological reason for the error*, has been precisely pinpointed (GOOD 1970/1971: p. 139).

Surprise indexes

A kind of alternative to the use of P -values are surprise indexes. The topic is reviewed by GOOD (1987b).

Probabilistic causality

Sometimes “causality” is taken to mean “determinism” as when people say that quantum mechanics sounded the death knell for causality. In the present context it is convenient to refer to determinism as *strict causality* and to refer to something less strict as *probabilistic causality*.

If the world is deterministic then probabilistic causality does not exist, but we will never know with certainty whether determinism or indeterminism is true. So it is legitimate to assume indeterminism even if it is only a convenient fiction, somewhat like using the axiom of choice in a mathematical proof. There would be no criminal law if, believing in determinism, we always said “*Tout comprendre c'est tout pardonner*”. Anyway *nous ne tout comprendrons jamais*, we never understand everything.

It is essential to make a distinction between the tendency of one event F to cause a later one E , denoted by $Q(E : F)$, and the extent to which F actually caused E , denoted by $\chi(E : F)$. In the law, a simple example is the distinction between murder and attempted murder. The distinction is important because the law rewards inefficiency in this case, at least in many countries.

The notations $Q(E : F)$ and $\chi(E : F)$ are both only abbreviations because one must allow also for the state U of the universe just before F occurred and also for all true laws of nature. It is also necessary to allow for the negations of E and of F but when you put all these aspects into the notation in a lecture some people walk out because they think you are doing mathematics.

It is extremely difficult to find a fully satisfactory explicatum for χ , though I think I have made some contribution towards it. (For my latest effort see my reply to a valid criticism by Salmon in Good (1987a).) Here I shall discuss only Q which seems to me to be of much greater importance in statistics, though in legal matters χ is at least as important. It will be Q that counts when you reach “*dem pearly gates*”. I will not say much even about Q because I have recently given two lectures on the topic (Good 1984/1985, 1987a).

The old-fashioned name “the probability of causes” referred to the application of Bayes’s theorem, where the “hypotheses” are regarded as mutually exclusive possible “causes” of some event or events. For example, the “event” might be a set of medical indicants and the possible “causes” might be various disease states. The topic I am discussing now is different: it refers to the tendency of some event F to cause another one E , not the probability that F was *the* cause of E .

Let us assume that $Q(E : F)$ is some function of all probabilities of the form $P(A | B)$ where A and B are logical combinations of E and F . This comes to the same thing as assuming that Q depends only on $P(E | F)$, $P(E | \bar{F})$ and $P(F)$. The probabilities are here best regarded as physical (propensities) because I am thinking of probabilistic causality as some-

thing that exists even if no conscious being is around. By assuming several desiderata related to the “causal strengths” and “causal resistances” of causal networks, we can arrive at the explicatum that Q is equal to the weight of evidence against F provided by the non-occurrence of E ; that is,

$$Q(E : F) = W(\bar{F} : \bar{E}) = \log \left[\frac{1 - P(E : \bar{F})}{1 - P(E : F)} \right] \quad (5)$$

(where of course U is taken for granted throughout). This expression is mathematically independent of $P(F)$ and this could have been taken as a desideratum though it was not explicitly used to obtain the explicatum. That $Q(E : F)$ is mathematically independent of $P(F)$, the initial probability of F , is desirable for the following reason.

In a scientific experiment we might decide whether to apply a treatment F by using a randomizing device that would determine $P(F)$. The purpose of the experiment might be to find out to what extent F causes E by repeating the experiment many times. It would be contrary to the spirit of scientific experimentation if the conclusion were to depend on our arbitrary choice of $P(F)$. Some people would go further and would say that no reliable conclusions are possible unless the experimenter uses a randomizing device to control whether F occurs. In this way we can be convinced that E and F did not have a common cause unless we believe in some possibly magical or paranormal effect that relates the randomizing device to the effectiveness of the treatment. This is why it is reassuring to discover that the proposed explicatum for Q does not depend on $P(F)$ although this property was not used in the original derivation of the explicatum.

It seems intuitively right that Q should have something to do with weight of evidence. So what happens if we define $Q(E : F)$ by some other weight of evidence, the possibilities being (i) $W(F : E)$, (ii) $W(E : F)$ and (iii) $W(\bar{E} : \bar{F})$? The second and third possibilities can be excluded because they depend on the initial probability of F , $P(F)$; so the only rival to $W(\bar{F} : \bar{E})$ is $W(F : E)$, still conditional on U of course. This rival will now be ruled out. Consider the “game” of Russian Roulette. In a self-explanatory notation, and for an obvious slightly oversimplified model, we have $P(E | F) = \frac{1}{6}$, $P(E | \bar{F}) = 0$, if the game is played with a six-shooter that contains just one bullet. Hence $W(\bar{F} : \bar{E}) = \log \left(\frac{6}{5} \right) = 78$ centibans (or “centicausits”), whereas $W(F : E) = \log \left[\frac{5}{6} / 0 \right] = \infty$. It makes sense that a necessary cause of E should have only a finite tendency to

cause E , while a sufficient cause should have an infinite tendency if E was not already inevitable. Playing Russian Roulette is a necessary but not a sufficient cause for disaster in the assumed model. Similarly, trying to cross the road is usually a necessary cause for getting run over, but fortunately it is not sufficient. Thus $W(F : E)$ is shot down. We see then that if $Q(E : F)$ is to be expressed in terms of weight of evidence there is really only one serious candidate, namely $W(\bar{F} : \bar{E})$. Moreover, this has desirable additive properties (e.g. Good 1983e: p. 209) that would not be shared by any function of it other than a mere multiple.

It turns out that $Q(E : F)$, as thus explicated, is identical with one of the measures of association used for 2 by 2 contingency tables. Also there is a relationship to the theory of linear regression (Good 1980, 1984/1985, 1987a).

Explicativity

POPPER (1959) suggested that a measure or index of explanatory power should be developed and this was a main theme of some of my later work, where I used the term *explicativity* (Good 1968b, 1977b).

By explicativity η is meant the extent to which one proposition or event explains why another should be believed, to express the matter a little too briefly. The concept is not intended to capture all the senses of "explanation". A desideratum-explicatum approach was used leading quickly to the explicatum

$$\eta(E : H) = \log P(E | H) - \log P(E) + \gamma \log P(H) \quad (6)$$

where $0 < \gamma < 1$, and where $\gamma = \frac{1}{2}$ might be adequate. We can think of γ as a clutter constant because the more we object to cluttering H with irrelevancies, the larger we would make γ .

The amount by which the explicativity of H exceeds that of H' is

$$\eta(E : H/H') = (1 - \gamma)W(H/H' : E) + \gamma \log O(H/H' | E), \quad (7)$$

a compromise between the weight of evidence provided by E on the one hand, and the posterior log-odds on the other hand. If we take $\gamma = 1$ there is no better hypothesis than a tautology such as $1 = 1$. If we take $\gamma = 0$ we ignore the prior probabilities. We must therefore compromise.

If explicativity is regarded as a kind of quasi-utility its maximization leads to a method for choosing among hypotheses, and this principle can be used in statistical problems of both estimation and significance testing. The results of applying this method make intuitive sense in several examples. For example, the method leads to interval estimation of parameters in these examples without assuming in advance that interval estimates should be used. The method is very general and could be used, for example, for the selection of regressor variables. The result would resemble methods proposed by AKAIKE (1974) and by SCHWARZ (1978). The maximization of *expected* explicativity is a reasonable recipe for experimental design, and it can be seen that γ then becomes irrelevant and the method reduces to that cited soon after (3).

The notion of explicativity seems appropriate for a semi-quantitative discussion of how good Natural Selection is as an explanatory theory as compared with other theories of evolution (GOOD 1986b).

Adhockery

When a hypothesis or theory H appears to be undermined by the total relevant evidence E , a defender of H might patch it up by changing it to a more elaborate hypothesis H' . Then has H been improved or is the change merely *ad hoc*? The concept of explicativity provides at least a formal solution to this problem: the change is *ad hoc* if $\eta(E : H/H')$ is positive, and $\eta(E : H/H')$ is a measure of the adhockery. If it is negative then the change is justified (cf. GOOD 1983c).

“Scientific method”

Somewhat supplementary to what I have said in this lecture is an encyclopedia article entitled “Scientific method and statistics” (GOOD 1985/1987). In that article I tried to define scientific method in terms of fourteen facets and to argue that statistics makes use of all of these facets. This does not show that statistics is identical with the scientific method but only that statistics is one example of the method. For each way of assigning weights to the facets one gets a different interpretation of “scientific method”.

Exploratory data analysis

At first it might seem that Exploratory Data Analysis is non-philosophical but I have argued in GOOD (1983d) that it has implicit Bayesian aspects.

Technique versus philosophy

Because I have been emphasizing the interface between philosophy and statistics, I might have given the impression that statistics is nothing but philosophy. That has not been my intention. Much of statistics consists of techniques for condensing data sets into simplified numerical and graphical forms that can be more readily apprehended by the eye-brain system, a system that has evolved at a cost of some 10^{18} organism-hours. Philosophers recognize the importance of techniques and technicians should reciprocate.

Acknowledgement

This work was supported in part by a grant from the National Institutes of Health.

References

- AKAIKE, H., 1974, *A new look at the statistical identification model*, IEEE Trans. Autom. Control 19, pp. 716–723.
- BLACK, M., 1967, *Induction*, in: The Encyclopedia of Philosophy, Vol. 4 (Macmillan and The Free Press, New York), pp. 169–181.
- CARD, W. I. and GOOD, I.J., 1974, *A logical analysis of medicine*, in: R. Passmore and J.S. Robson, eds., *A Companion to Medical Studies* (Blackwell, Oxford), Vol. 3, Chap. 60.
- CARNAP, R., 1950, *Logical Foundations of Probability* (University of Chicago Press, Chicago, IL).
- CHRISTENSEN, R., 1983, *Multivariate Statistical Modelling* (Entropy Limited, Lincoln, MA).
- DE FINETTI, B., 1937/1964, *Foresight: its logical laws, its subjective sources*, trans. from the French of 1937 by H. Kyburg, in: H.E. Kyburg and H.E. Smokler, eds., *Studies in Subjective Probability* (Wiley, New York), pp. 95–158, corrected in 2nd edn., 1980, pp. 55–118.
- FINE, T.L., 1973, *Theories of Probability* (Academic Press, New York).
- FISHER, R.A., 1955, *Statistical methods and scientific induction*, J.R. Stat. Soc. B 17, pp. 69–78.
- GIBBS, J.W., 1875/1906/1961, *On the equilibrium of heterogeneous substances*, in: The

- Scientific Papers of J. Willard Gibbs, Vol. 1 (Longmans, Green & Co., London), reprinted 1961 by Dover Publications, New York, pp. 55–349.
- Good, I.J., 1950, *Probability and the Weighing of Evidence* (Charles Griffin, London and Hafners, New York).
- Good, I.J., 1952, *Rational decisions*, J.R. Stat. Soc. B, 14, pp. 107–114. Reprinted in Good Thinking, 1983e.
- Good, I.J., 1955/1956, *Some terminology and notation in information theory*, Proc. Inst. Elect. Eng. C, 103, pp. 200–204; or 1955, Monograph 155R.
- Good, I.J., 1957, *Saddle-point methods for the multinomial distribution*, Ann. Math. Stat. 28, pp. 861–881.
- Good, I.J., 1958, *Significance tests in parallel and in series*, J. Am. Stat. Assoc. 53, pp. 799–813.
- Good, I.J., 1959, *Kinds of probability*, Science 129, pp. 443–47. Reprinted in Good Thinking 1983e.
- Good, I.J., 1961/1962, *A causal calculus*, Br. J. Philos. Sci. 11, pp. 305–318; 12, pp. 43–51; 13, p. 88. Reprinted in Good Thinking, 1983e.
- Good, I.J., 1963, *Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables*, Ann. Math. Stat. 34, pp. 911–934.
- Good, I.J., 1965, *The Estimation of Probabilities: An Essay on Modern Bayesian Methods* (MIT Press, Cambridge, MA).
- Good, I.J., 1966, *How to estimate probabilities*, J. Inst. Math. Applic. 2, pp. 364–383.
- Good, I.J., 1967, *A Bayesian significance test for multinomial distributions*, J.R. Stat. Soc. B 29, pp. 399–431 (with discussion).
- Good, I.J., 1968a, *Some statistical methods in machine-intelligence research*, Virginia J. Sci. 19, pp. 101–110.
- Good, I.J., 1968b, *Corroboration, explanation, evolving probability, simplicity, and a sharpened razor*, Br. J. Philos. Sci. 19, pp. 123–143.
- Good, I.J., 1969, *What is the use of a distribution?*, in: P.R. Krishnaiah, ed., *Multivariate Analysis II* (Academic Press, New York), pp. 183–203.
- Good, I.J., 1970/1971, *The probabilistic explication of information, evidence, surprise, causality, explanation, and utility*, in: V.P. Godambe and D.A. Sprott, eds., *Foundations of Statistical Inference: Proc. Symp. on the Foundations of Statistical Inference held at the University of Waterloo, Ontario, Canada, from March 31 to April 9, 1970* (Holt, Toronto), pp. 108–141 (including discussion).
- Good, I.J., 1977a, *Dynamic probability, computer chess, and the measurement of knowledge*, in: E.W. Elcock and D. Michie, eds., *Machine Intelligence 8* (Ellis Horwood Ltd. and Wiley), pp. 139–150. Reprinted in Good Thinking, 1983e.
- Good, I.J., 1977b, *Explicativity: a mathematical theory of explanation with statistical applications*, Proc. R. Soc. (London) A 354, pp. 303–330; 1981, *erratum* 377, p. 504. Reprinted in Good Thinking, 1983e.
- Good, I.J., 1979/1981, *Some history of the hierarchical Bayesian methodology*, in: *Trabajos de Estadística y de Investigación Operativa*. Also in: J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain), May 28 to June 2, 1979*, University of Valencia, pp. 489–510, 512–519 (with discussion). Reprinted in Good Thinking, 1983e.
- Good, I.J., 1980, *Degrees of causation in regression analysis*, C71 in *J. Stat. Comput. Simulation* 11, pp. 153–155.
- Good, I.J., 1981a, *Some logic and history of hypothesis testing*, in: J.C. Pitt, ed., *Philosophy in Economics*, University of Western Ontario series on the Philosophy of Science (D. Reidel, Dordrecht), pp. 149–174. Reprinted in Good Thinking, 1983e.
- Good, I.J., 1981b, *Can scientific induction be meaningfully questioned?*, C100 in *J. Stat. Comput. Simulation* 13, p. 154.

- GOOD, I.J., 1982, *Standardized tail-area probabilities*, C140 in *J. Stat. Comput. Simulation* 16, pp. 65–66.
- GOOD, I.J., 1983a, *The robustness of a hierarchical model for multinomials and contingency tables*, in: G.E.P. Box, T. Leonard and C.-F. Wu, eds., *Scientific Inference, Data Analysis, and Robustness* (Academic Press, New York), pp. 191–211.
- GOOD, I.J., 1983b, *Scientific induction: universal and predictive*, C143 in *J. Stat. Comput. Simulation* 16, pp. 311–312.
- GOOD, I.J., 1983c, *A measure of adhocery*, C145 in *J. Stat. Comput. Simulation* 16, p. 314.
- GOOD, I.J., 1983d, *The philosophy of exploratory data analysis*, *Philos. Sci.* 50, pp. 283–295.
- GOOD, I.J., 1983e, *Good Thinking: The Foundations of Probability and its Applications* (University of Minnesota Press, Minneapolis).
- GOOD, I.J., 1983/1985, *Weight of evidence: a brief survey*, in: J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting* (September 6/10, 1983) (North-Holland, New York), pp. 249–269 (including discussion).
- GOOD, I.J., 1984a, *How should tail-area probabilities be standardized for sample size in unpaired comparisons?*, C191 in *J. Stat. Comput. Simulation* 19, p. 174.
- GOOD, I.J., 1984b, *The best explicatum for weight of evidence*, C197 in *J. Stat. Comput. Simulation* 19, pp. 294–299; 20, p. 89.
- GOOD, I.J., 1984c, *Standardized tail-area probabilities: a possible misinterpretation*, C199 in *J. Stat. Comput. Simulation* 19, p. 300.
- GOOD, I.J., 1984d, *The tolerant Bayesian's interpretation of a tail-area probability*, C200 in *J. Stat. Comput. Simulation* 19, pp. 300–302.
- GOOD, I.J., 1984/1985, *Causal propensity: a review*, in: P.D. Asquith and P. Kitcher, eds., *PSA 1984, Vol. 2* (Philosophy of Science Association), pp. 829–850.
- GOOD, I.J., 1985/1987, *Statistical method and statistics*, *Encycl. Stat. Sci.* 8, pp. 291–304.
- GOOD, I.J., 1986a, *Some statistical applications of Poisson's work*, *Stat. Sci.* 1, pp. 157–170 (with discussion).
- GOOD, I.J., 1986b, *"Neo"-Darwinism*", *Physica* 22D, pp. 13–30. Also in: D. Farmer, A. Lapedes, N. Packard, and B. Wendroff, eds., *Evolution, Games, and Learning* (North-Holland, Amsterdam), pp. 13–30.
- GOOD, I.J., 1986c, *The whole truth*, *Inst. Math. Stat. Bull.* 15, pp. 366–373. (An editor's invited column.)
- GOOD, I.J., 1986d, *The computer-intensive form of a Bayes/non-Bayes compromise*, C265 in *J. Stat. Comput. Simulation* 26, pp. 132–133.
- GOOD, I.J., 1987a, *Causal tendency: a review*, a revision of Good (1984/1985) as an invited paper at the *Conference on Probability and Causation at Irvine, California, 1985, July 15–19*. In: W. Harper and B. Skyrms, eds., *Causation, Chance and Credence* (Reidel, Dordrecht), pp. 23–50, 73–78.
- GOOD, I.J., 1987b, *Surprise index*, *Encycl. Stat. Sci.* 9, pp. 104–109.
- GOOD, I.J. and CARD, W.I., 1971, *The diagnostic process with special reference to errors*, *Methods Inform. Med.* 10, pp. 176–188.
- GOOD, I.J. and CROOK, J.F., 1974, *The Bayes/non-Bayes compromise and the multinomial distribution*, *J. Am. Stat. Assoc.* 69, pp. 711–720.
- GOOD, I.J. and DEATON, M.L., 1981, *Recent advances in bump-hunting*, in: W.F. Eddy, ed., *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (Springer, Berlin) pp. 92–104 (with discussion).
- GOOD, I.J. and GASKINS, R.A., 1971, *Non-parametric roughness penalties for probability densities*, *Biometrika* 58, pp. 255–277.
- GOOD, I.J. and GASKINS, R.A., 1972, *Global nonparametric estimation of probability densities*, *Virginia J. Sci.* 23, pp. 171–193.

ASTRONOMICAL IMPROBABILITY

IAN HACKING

*Institute for the History and Philosophy of Science and Technology, University of Toronto,
Victoria College, Toronto, Canada M5S 1K7*

This paper is about probability in astronomy and astrophysics, but the title has another connotation as well. “Astronomical” means immensely large. Any enormous number or proportion can be called astronomical, and, in the case of a state lottery, we can well say that the odds against winning are astronomical. I shall be concerned, in part, with the astronomical improbabilities that occur in astronomy. The odds in state lotteries are rather favourable compared to the odds from time to time adduced in cosmological speculations. We find these odds almost at the beginning of probability mathematics. Thus BERNOULLI (1734) wrote a prize essay on the question: is there a reason why the planets move around the sun in roughly coplaner orbits? He constructed three different models, and found that the odds against this arising from chance are 17^5 to 1, 13^5 to 1 and 12^6 to 1. The first odds are comparable to a lottery with 7 billion tickets.

My examples will be more current than that. One section draws on gravitational lensing, a topic that has obtained observational content only in the past decade. Another treats of the anthropic principle in cosmology, about 15 years old. I also examine the probabilistic arguments used to show that we live in an inhomogeneous “clumpy” universe, very much a matter of contemporary investigation, but also one of the oldest topics in probability theory, going back to MICHELL (1767). Here, on occasion, we shall find truly impressive odds being proffered. Would you take 10^{40} to 1 against the hypothesis that clusters of galaxies in the very distant part of the universe are distributed at random? (ABELL 1958).

One may find something inherently dubious in very large improbabilities adduced as grounds for belief or disbelief. (Of course I have no complaint about minute probabilities that occur in abstract models of physical systems, as in statistical mechanics; I am here speaking of the

epistemic use of probability to infer conclusions on the basis of available evidence.) There is only one field of enquiry that regularly advances improbabilities as gross as those sometimes adduced in astrophysics: psychical research. The latter has been generating vast improbabilities and corresponding certainties for just over a century. Writing of one remarkable early work (GURNEY *et al.* 1886), PEIRCE (1887) wrote that the authors “cipher out some very enormous odds in favor of the hypothesis of ghosts. I shall not recite these numbers, which captivate the ignorant, but which repel thinking men, who know that no human certitude reaches such figures of trillions, or even billions, to one.” We might speak of Peirce’s maxim: attach little epistemic weight to enormous improbabilities. If you are moved by great improbabilities, remember psychical research. That is enough of a sceptical preamble. The paper comes in three parts: (a) Ordinary uses of “improbability reasoning” in astrophysics. There are at least three distinct such uses. It is convenient to see how each may occur in the same field of enquiry, and for this purpose I take gravitational lensing. (b) An extraordinary use of “improbability reasoning”: discussions of the clumpiness of the universe. (c) Reasoning by hyperimprobability: the anthropic principle.

Ordinary uses of “improbability reasoning” in astrophysics

Gravitational lensing was predicted as early as 1919 by the Newtonian, Oliver Lodge, and gravitational lens effects have been observed since 1979. HACKING (1989) provides an account of observation, theory and history accessible to philosophers, together with references to the technical literature. The basic idea is this. Suppose that there is an object at a great distance from us: in recent examples, a quasi-stellar object (QSO). Suppose further that there is a massive object almost on the line of sight between us and the QSO. Now radiation (be it in the radio or optical spectrum) from the QSO, grazing either side of the heavy object, will be deflected towards that object. Thus under suitable conditions we would expect to see two images of the QSO, one from the radiation deflected towards us on one side of the intervening object, and the other deflected towards us on the other side. (This is the first simple analysis, postulating that the intervening object is a point mass; it is readily shown that for spatially distributed masses there should be an odd number of images.) If the heavy object were exactly on the line of sight between us and the QSO, we would expect to see a halo around the heavy object. The

astrophysics is vastly more complex than this simple model suggests, but this sketch will suffice for my purposes. I shall now state three distinct “ordinary” uses of improbability in reasoning about gravitational lensing.

(i) In judging that two or more images are the product of gravitational lensing. The first “well-confirmed” gravitational lens was published by WALSH *et al.* (1979) under the title, “0957 + 651 A, B: Twin quasistellar objects or gravitational lens?” The title is self-explanatory. There are two images A, B. Are they images of twinned QSOs, or of a single QSO imaged by lensing? Images A and B are astonishingly similar, for example in the emission and absorption spectra. One considers the angular separation of the images, their shapes and sizes in the lens model. If one uses the magnification of A compared to B (in the lens model), one deduces a luminosity characteristic of a QSO.

Intuitively, the similarity between A and B is too great to attribute to coincidence. This reasoning can be formally cast as a Fisherian significance test. Either, by chance, there are two incredibly similar QSOs in the same region of the sky, or else we have a gravitational lens. Probabilities can be computed, and the null hypothesis rejected. (There is of course another possibility, that we have twin QSOs produced by a single mechanism, but this is dismissed for the present because one has not the slightest idea what such a mechanism could be.) As always one can provide a Bayesian re-analysis, considering the prior probability of gravitational lensing, computing the likelihoods of the coincidence on that hypothesis and on the hypothesis of random QSO distribution, and obtaining a posterior probability.

Astrophysicists are in fact loth to quantify. “We do not think that a useful a posteriori statistical test of this assertion [of lensing] can be carried out” (WALSH *et al.* 1979: p. 383). Others are ironic: “We have not attempted to calculate a posteriori a probability for this event; but for those who would do so. . .” (YOUNG *et al.* 1980: p. 519). Or they are nervous and apologetic: “Although estimation of probability after the fact is often a dangerous exercise, it can be informative” (HUCHRA *et al.* 1985: p. 693). Would that philosophers of probability were as aware of the difficulties of drawing probable inferences from real data! It is evident, however, that the methodologies to be employed here are exactly the same as those that occur in mundane science.

(ii) Infrequencies that call in question the entire theory. Statistical considerations also enter in a quite different way. On every account of lensing (except a calculation of EINSTEIN (1933)!) detectable gravitational lenses should not be very uncommon. ZWICKY (1937) predicted one

detectable lens system per 100 nebulae. Speaking of "passages" (of one object in front of another to produce a lens), REFSDAL (1964) wrote that "it seems safe to conclude that passages observable from the earth occur rather frequently." WEEDMAN *et al.* (1982) remark that "It is emphasized that survey techniques such as that used to discover this pair of images found over 1000 quasars, and we are puzzled as to why pairs of smaller separation have not been found in this way." And so it goes. There is a discrepancy between the frequency of observed lenses, and the probability with which they "ought" to occur. Something must be wrong with either the theory, or some assumptions about the conditions of its application. Here, as with (i), we have an example of improbability reasoning that can be found in any branch of science.

(iii) Lensing as a mass statistical phenomenon. The distinction between "two concepts of probability" has been too often put as a difference between an epistemic concept on the one hand (subjective, personal, logical, or whatever) and a physical concept (frequency, tendency, disposition, propensity) on the other. The core difference is between probability as a tool in inference and probability as a tool in modelling. (i) and (ii) above concern inference. Inevitably, one can also make probability models of lenses. Thus far we have been thinking of one distant object imaged by one intervening heavy object. But there are myriad objects standing in the way of radiation from a QSO, and it is proposed that we have a system of "microlenses" that can only be studied statistically.

The consequences of microlensing could be astonishing. For example, there is a class of extremely variable polarized extragalactic sources with significant emission all the way from the radio spectrum to X-rays: the BL Lac objects of which 87 occur in an appendix to the 1987 QSO catalogue. OSTRIKER and VIETRI (1985) propose that at least some of the "high variability" that characterises these is a consequence of passage of large numbers of bodies between us and the objects, causing very variable minilensing. The BL Lacs may be QSOs after all.

More generally the "evolution" of quasi-stellar objects is of great current interest. Could this evolution be an artifact of the passage of many microlenses, changing the images of the QSO? (TURNER 1980; PEACOCK 1982). Or consider this: QSOs are more common near galaxies. Could this be a consequence of microlensing by the galaxies, that makes the QSOs more detectable there? (SCHNEIDER 1986). And so on. Such questions can only be treated statistically, although the problems are horrendous for a real, i.e. clumpy, universe. One can at best hope for mathematically coherent statistical models; to take the title of a recent

study, “Self-consistent probabilities for gravitational lensing in inhomogeneous universes” (EHLERS and SCHNEIDER 1986).

Uses (i) and (ii) of probability had to do with inference; (iii) has to do with modelling. Despite the difficulties of modelling in this third use, any philosophical issues that arise are logically identical to those that occur in at least the other natural sciences.

An extraordinary use of improbability: do we live in a clumpy universe?

MICHELL (1767) thought about the fact that there are a good many pairs, triplets or groups of stars close together. There are 230 stars equal in magnitude to the pair Beta Capricorni. What is the probability that if 230 points are distributed at random on the sphere, two will be as close together as Beta Capricorni? Only $1/80$, he answers. The Pleiades is a group of six stars. There are 1500 stars of comparable magnitude. “We shall find the odds to be near 500,000 to 1, that no fixed stars, out of that number, scattered at random, in the whole heavens, would be so small a distance from each other, as the Pleiades are.” He considers whether the contiguity of the Pleiades might be due to “their mutual gravitation or some other law or appointment of the Creator”. Whatever it is, he concludes that the clustering of stars shows that they are “under the influence of some general law”.

FISHER (1956: p. 39) recomputed and found lesser odds, namely 33,000 to 1. He took the reasoning to be that of a significance test. There is a null hypothesis, that the stars are distributed at random. The fraction $1/33,000$, wrote Fisher, “is amply low enough to exclude at a high level of significance any theory involving a random distribution”. One reason I speak of “extraordinary” improbabilities in this section is that Fisher taught us to think in terms of significance levels of 1% or 5%. In Fisher’s recomputed version of Michell’s problem, we are talking about 0.003%. It is far from clear whether this is a difference in degree, or positively a difference in kind.

Numerous students of foundations have reflected on the logic of Michell’s problem. Fisher himself refers to BOOLE (1854); the pair of them thought that no Bayesian analysis of Michellian reasoning could be other than *ad hoc*. (Boole also discussed the Bernoullian problem mentioned above, of the cause of the approximate coplanarity of the planetary orbits.) Undoubtedly the greatest influence of Michell’s essay was upon William Herschel. His allusions to and acknowledgements of

Michell occur chiefly in **HERSCHEL** (1782), which includes a vast catalogue of double stars, augmented in **HERSCHEL** (1785). Surveying the heavens was a family occupation, William's only peers being his sister Caroline and his son John. The latter, well aided by his sons, nephews and nieces, produced the *General Catalogue of Nebulae and Clusters of Stars* (**HERSCHEL** 1864), which in turn notes the remarkably clumpy character of the universe. A rather full account of this "Herschelian" epoch in the study of our inhomogeneous universe is provided by **LANDMARK** (1927).

Edwin Hubble changed all that. He was the first to establish that our Galaxy is only one among countless galaxies. The Galaxy is certainly inhomogeneous, but let us expand our vision. Galaxies may be local clumps, but may not a view of the universe upon a large enough scale reveal an essential uniformity? His first step was a painstaking justification of his assumption that "the stars involved in the nebulae are directly comparable with the stars in our own system" (**HUBBLE** 1926: p. 356). He became confident that "for the first time, the region now observable with existing telescopes may possibly be a fair sample of the universe as a whole" (**HUBBLE** 1934: p. 8). Then: "any considerable collection of nebulae, chosen at random, should be a fair sample" of the universe (p. 57). In his book of lectures (**HUBBLE** 1936) "the observable region is not only isotropic but homogeneous as well" (p. 31). "Now the observable region is our sample of the universe. . . . If the sample is fair, its observed characteristics will determine the physical nature of the universe as a whole. . . . And the sample may be fair. . . . Thus for purposes of speculation we may apply the principle of uniformity and suppose that any other portion of the universe, selected at random, is much the same as the observable region" (pp. 34–35).

There are two elements to these assertions. First there is the properly cautious "purposes of speculation", the "may be fair". Then there is the flat assertion that the observable universe is homogeneous and isotropic. How does that square with the opinion of the Herschel family? It was sound for their time, one could reply, but now we can sweep a vastly larger portion of the universe and, Hubble asserts, we find that the Herschelian clusters are merely local phenomena. This assertion, which sounds like plain fact from the pen of so careful an observer, was in fact more like an article of faith, the metaphysical theme that the universe must, in the large, be uniform.

This doctrine became elevated to a principle and given a name worthy of its scope: the "cosmological principle" that "the universe presents the same aspect from every point except for local irregularities" (**BONDI** 1952:

p. 11). Then there is the “perfect cosmological principle” to the effect that “apart from local irregularities the universe presents the same aspect from any place at any time” (p. 12).

A methodologist of astronomy put the most direct challenge to this grand vision of nature. DINGLE (1953) observed that we all agree to the principle (in his italics) that “*no statement about the universe, or nature, or experience, or whatever term you prefer for the object of scientific investigation, shall be made*—let alone advanced as a fundamental principle—*for which there is no evidence*. What we are faced with now is the quite different claim that *any statement may be made about it that cannot immediately be refuted*” (p. 396).

Methodological irony does not compel the adventurous; it was the 3°K of PENZIAS and WILSON (1965) that put the Big Bang theory back on its feet and hence put paid to the “perfect cosmological principle”. Yet the fact that this background microwave radiation is remarkably isotropic (PARTRIDGE and WILKINSON 1967) on a largish scale was consistent with Hubble’s cosmological principle.

How do we test the proposition that even the observable region of the universe is uniform? This problem attracted the dean of statisticians (NEYMAN *et al.* 1953, 1956). But such was the *metaphysical* pull of the belief that we live in a uniform universe, that singularly little was done after Hubble to test the belief seriously. There were a few voices, and DE VAUCOULEURS (1971) was able to review “the observations and statistical evidence now available which prove conclusively that clumpiness (i.e. clustering) is a basic property of the distribution of galaxies on all observable scale” (p. 114). I need hardly say that this talk of “proof” was rhetorical; few agreed with him at the time.

From a logical point of view there is little significant statistical advance from MICHELL (1767) to the main body of work reported by DE VAUCOULEURS (1971). “Let a survey area A including a total of N_T points be divided into m equal fields. . . if the points are scattered at random the number $N(n)$ of fields including n points is given by the binomial law. . . if the number of fields is large, i.e. if $p \ll 1$, the binomial distribution is approximated by. . . the Poisson distribution. . . when n is large the Poisson distribution tends towards the normal (symmetric) distribution. . . Hence a simple way to test an observed frequency distribution. . . is to compute the ratio of the observed dispersion. . . to the theoretical value.” That, with minor variations (and thoughtful interventions by e.g. Neyman) is what people did. The sky or a portion of the sky was divided into cells and then some sort of parametric or at any rate normal-

distribution based test such as chi-squared was made on the “null hypothesis” of randomness.

ABELL (1958) used such a form of reasoning to investigate the idea that there are superclusters of galaxies, i.e. clusters of clusters of galaxies. He concluded that the odds against random distribution of clusters, at very great distances, were between 10^{20} and 10^{40} to 1 — odds to which I am glad to apply Peirce’s maxim, i.e. pay no attention to them. YU and PEEBLES (1969) describe several grounds for complaint which are of great interest to the student of statistical inference. I shall mention only one.

How does one choose the cell size? This is always a question, but it is important to look at the data to see if even before analysis there is a potential for systematic error. Censoring is now an all too familiar problem in observational astronomy. For all sorts of reasons connected with the heavens themselves we see some bits better than others, or we see some objects better for reasons that have nothing to do with the objects (cf. the love of quasi-stellar objects for galaxies mentioned in (iii) in the previous section above). It happens that the mean surface density of clusters at galactic latitudes of around 60° or 70° is almost double that at the extreme higher or lower latitudes. It is vastly more plausible that this is an artifact of some kind of censoring, than that it has a physical basis in the clusters themselves. It is then to be expected that a chi-squared test (or whatever) applied to equal cells will show strong significance. The effect is heightened by Abell’s using as cell size the cell giving the smallest probability to a supercluster — heightened to 10^{20} or more. More generally the problem is one of choosing the scale according to which one will recognise inhomogeneity: cell size, in simplified shorthand.

Yu and Peebles turned to the power spectrum analysis that had been devised by Tukey (BLACKMAN and TUKEY 1959). Among its numerous virtues is that it provides a ready computation of the extent of irregularity on any chosen scale. Moreover the analysis is not only rather insensitive to censoring effects, but also there are ways to estimate the degree to which various effects show up in the apparent power spectrum. An immediate consequence of this work was the conclusion that there is not yet strong statistical evidence for Abell’s hypothesis of superclusters.

PEEBLES (1973) states the general theory of power spectrum analysis as applied to astronomical catalogues, and in a subsequent series of a dozen papers ranging over a decade, he and his students have systematically noted the evidence for and against different kinds of inhomogeneity in the observable universe. That this has been something of a growth industry may be seen from conference proceedings (MARDIROSIAN 1984).

A number of outstanding logical problems are suggested in several papers at that conference, particularly GELLER (1984).

All the work from Michell to Peebles is limited in an obvious way. A clumpy universe is three-dimensional. Past statistical surveys occur on the surface of the celestial sphere and are in effect two-dimensional. There simply has not been a systematic three-dimensional catalogue. Such catalogues by red shift are now in preparation. As substantial parts become complete (this means now), the entire statistical game is changed. However, this does not mean that we will automatically be able to answer questions about the statistics of clumpiness. A whole battery of new techniques will be evolved. Students of the foundations of statistical inference are urged to keep an eye on developments. If there is a field for innovation in "inductive logic" that is wide open at present, this is it.

Hyperimprobability: the anthropic principle

The anthropic principle was proposed and named by CARTER (1974). Antecedents are claimed at least as far back as BOLTZMANN (1897). It has received a good deal of attention from a small number of distinguished cosmologists and a few philosophers. BARROW (1986) is a thick and important source of information, argument and references. My one contribution to the debate has already been published in HACKING (1987a, b). Here I wish only to remark some odd features of the role of probability in the anthropic principles.

What is in question? Carter wrote of the "*anthropic principle*, to the effect that what we can expect to observe must be restricted by the conditions necessary for our presence as observers". He also introduced "what may be termed the *strong anthropic principle* stating that the universe (and hence the fundamental parameters on which it depends) must be such as to admit the creation of observers at some stage". The latter is a remarkable proposition that combines metaphysical appeal with what appears to be a strong claim to matters of fact. I shall say nothing about it, except perhaps to murmur the methodological scepticism of Herbert Dingle quoted above. The simple anthropic principle in contrast looks bland enough to be a tautology. Why state it at all?

At issue is the apparent "fine tuning" of the universe, of which DAVIES (1982, 1983) gives a favourable account for the general reader. ROBSON (1987) is the proceedings of a conference (prompted in part by Davies'

popular expositions) in which experts from various fields discuss whether there is fine tuning, and if so, so what.

What is meant by fine tuning? Suppose that the fundamental laws of nature are fixed, but with free parameters for the fundamental constants, and with boundary conditions that are flexible. Thus the velocity of light, the Hubble constant, and the ratio of protons to antiprotons just after the big bang are alike undetermined. Now this notion of given laws of nature, prior to anything, is metaphysically grand, but philosophically suspect (HACKING 1987a: pp. 127–130). But let us take it for granted. We note that under most assignments of value to parameters and under most boundary conditions, the universe is unstable. It collapses, blows up or whatever. It takes a lot of fixing to get some enduring matter, let alone galaxies, let alone the Galaxy, let alone the solar system, let alone earth, let alone life, let alone observers. Each existent item in this list requires more and more fine tuning of parameters and initial conditions.

Such reflections have provoked, in some minds, a revival of the argument from design for the existence of an intelligent, designing creator. An argument, in which the anthropic principle is embedded, is taken to counter this. COLLINS and HAWKING (1973) answered their question “Why is the universe isotropic?” in this vein: “Because we are here.” Less briefly, we should not be surprised that the universe has something like the structure we find in it, because only a universe thus structured could have observers in it.

I must here interrupt the argument to note that design is not what makes anthropic principles interesting for cosmology. Here is one serious use of the idea: many features of the universe, such as the distributions of mass, or the ratios of various fundamental particles, cannot be derived from any fundamental laws. These seem to be fortuitous features of the universe, but we can ask, what tuning was necessary to bring just these apparently accidental aspects of the world into existence? Answering, we gain some insight into structures of the universe that do not arise from the fundamental laws, but which are essential that our universe should exist.

To return to anthropism versus the designer god, the deist is not impressed by Hawking’s answer, yet. “Because we are here!” — “well, why are we here?” Collins and Hawking constructed a topology and probability measure over all possible universes consistent with the fundamental laws of nature and a big bang hypothesis. They found that in this topology the set of parameter assignments and initial conditions, consistent even with isotropy, is of measure zero. That is, there is a 0 probability that anything like our universe should exist: I call this

hyperimprobability. Hawking's derivation has been challenged: I mention it only to connect hyperimprobability and the anthropic principle.

"Why are we here?" asks the deist, "If observers are impossible in 'almost all' (measure-theoretically speaking) universes, does it not require a designer to bring our world into being?" There are two distinct anthropic moves made in reply. They are commonly taken to be essentially the same, but from a logical point of view they are entirely distinct. I label them the Carter move and the Wheeler move. The Carter move is wild metaphysics, makes no use of probability, and does confute the design argument. The Wheeler move is far less astounding, does use probability, and does not confute the deist.

Carter postulates that all possible universes consistent with the fundamental laws of nature and the big bang actually co-exist. Hence our universe exists, and no designer is called in to select just this universe (and it is a sheer tautology that we are observing an observable universe). The premise is mind-boggling, but the argument is valid. It is purely deductive. It makes no use of probability whatsoever.

The Wheeler move thinks not of co-existent universes but of sequential ones. Most universes are unstable. One comes into being, explodes, disappears. Another arrives, and then collapses into nothing. And so on, one at a time. The initial conditions and fundamental constants of each universe are thought of as random variables, and "almost all" lack sentient beings. But in the long sequence of universes, an occasional universe populated with observers comes into being. As a matter of logic, we are in one of these.

This style of reflection is familiar from the 17th century argument from design for the existence of God. Opponents held that if our one universe has been around long enough, then by chance the particles will sooner or later adopt their present configuration, including us, so chance, and a long enough time span, suffices to explain our existence. The Wheeler story merely adds a sequence of universes, and holds that in a long enough sequence, there is a probability of 1 that a universe finely tuned in our direction should come into being.

This reasoning is fallacious as an explanation of why our 0 probability universe should exist. It commits the inverse gambler's fallacy (HACKING 1987b). The gambler's fallacy is the error of thinking that if on a fair roulette there has been a long run of red, then black is more probable at the next spin in order to "even out" the outcomes. In the inverse gambler's fallacy, one observes an unusual event — say 4 dice are rolled, and each falls 6 up. How come? Well, says the inverse gambler, this rare

event is evidence that many throws must have occurred, but of course, if many throws have occurred, then it is hardly surprising that a quadruple six should occur. This is as fallacious as the gambler's fallacy. HACKING (1987b) analyses the fallacy in some detail and shows that it is formally identical to the Wheeler move used to rebut the creationist. We should not conclude that the creationist is right. We conclude only that postulating a long sequence of universes does no good in refuting the creationist. Instead we say to him simply: yes, an exceptionally improbable event has occurred.

Weak and strong anthropic principles can play certain suggestive roles in cosmology, but hyperimprobability plays no part in their legitimate use.

References

- ABELL, G.O., 1958, *The distribution of rich clusters of galaxies*. *Astrophys. J. Suppl.* 3, pp. 211–288.
- BARROW, J. and TIPLER, F.J., 1986, *The Anthropic Cosmological Principle* (Clarendon Press, Oxford).
- BERNOULLI, D., 1734, in: *Recueil des pièces qui ont remporté le prix de l'Académie Royale des Sciences* 3, pp. 95–122.
- BLACKMAN, R.B. and TUKEY, J.W., 1959, *The Measurement of Power Spectra* (Dover, New York).
- BOLTZMANN, L., 1897, *Hrn. Zermelos Abhandlung, 'Über die mechanische Erklärung irreversibler Vorgänge'*. *Ann. Phys.* 76, pp. 392–398.
- BONDI, H., 1952, *Cosmology* (Cambridge University Press, Cambridge).
- BOOLE, G., 1854, *An Investigation of the Laws of Thought, on which are Founded the Mathematical Theories of Logic and Probability* (London).
- CARTER, B., 1974, *Large number coincidences and the anthropic principles in cosmology. Confrontation of Cosmological Theories with Observational Data*, in: M.S. Longair, ed., *Proceedings of the 2nd Copernicus Symposium, 1973* (Reidel, Dordrecht).
- COLLINS, C.B. and HAWKING, S., 1973, *Why is the universe isotropic?* *Astrophys. J.* 180, pp. 317–334.
- DAVIES, P.C.W., 1982, *The Accidental Universe* (Cambridge University Press, Cambridge).
- DAVIES, P.C.W., 1983, *God and the New Physics* (Dent, London).
- DINGLE, H., 1953, *The President's Address*. *Mon. Not. R. Astron. Soc.* 113, pp. 393–407.
- EHLERS, J. and SCHNEIDER, P., 1986, *Self-consistent probabilities for gravitational lensing in inhomogeneous universes*. *Astron. Astrophys.* 168, pp. 57–61.
- EINSTEIN, A., 1936, *Lens-like action of a star by the deviation of light in the gravitational field*, *Science* 84, 506–507.
- FISHER, R.A., 1956, *Statistical Methods and Scientific Inference* (Oliver and Boyd, Edinburgh).
- GURNEY, E., MYERS, F.H. and PODMORE, F., 1886, *Phantasms of the Living* (Trubner, London).

- HACKING, I., 1987a, *Coincidences: mundane and cosmological*, in: ROBSON, 1987, pp. 119–138.
- HACKING, I., 1987b, *The inverse gambler's fallacy: the argument from design. The anthropic principle applied to Wheeler universes*. *Mind* 96, pp. 331–340.
- HACKING, I., 1989, *Extragalactic reality: the case of gravitational lensing*. *Philos. Sci.*, in press.
- HERSCHEL, J., 1864, *General catalogue of nebulae and clusters of stars*. *Phil. Trans. R. Soc.* 154, pp. 1–137.
- HERSCHEL, W., 1782, *On the parallax of the fixed stars*. *Phil. Trans. R. Soc.* 72, pp. 82–111. *Catalogue of double stars, ibid.*, pp. 112–162.
- HERSCHEL, W., 1785, *Catalogue of double stars*. *Phil. Trans. R. Soc.* 75 (i), pp. 40–126.
- HUBBLE, E., 1926, *Extra-galactic nebulae*. *Astrophys. J.* 64, pp. 321–369.
- HUBBLE, E., 1934, *The distribution of extragalactic nebulae*. *Astrophys. J.* 79, pp. 8–76.
- HUBBLE, E., 1936, *The Realm of the Universe* (Yale University Press, New Haven).
- HUCHRA, J. et al., 1985, 2237 + 0303: *A new and unusual gravitational lens*. *Astron. J.* 90, pp. 691–697.
- LANDMARK, K., 1927, *Studies of anagalactic nebulae*. *Uppsala Astr. Obs. Med. no.* 30.
- MICHELL, J., 1767, *An inquiry into the probable parallax, and magnitude of the fixed stars, from the quantity of light which they afford us, and the particular circumstances of their situation*. *Phil. Trans. R. Soc.* 57, pp. 234–264.
- NEYMAN, J. et al., 1953, *On the spatial distribution of galaxies: a specific model*. *Astrophys. J.* 117, pp. 92–133.
- NEYMAN, J. et al., 1956, *Statistics of images and galaxies with particular references to clustering*. *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability* 3, pp. 75–112.
- OSTRIKER, J.P. and VIETRI, M., 1985, *Are some BL Lac objects artefacts of gravitational lensing?* *Nature* 318, pp. 446–448.
- PARTRIDGE, R.B. and WILKINSON, D.T., 1967, *Large scale density inhomogeneities in the universe*. *Nature* 215, p. 719.
- PEACOCK, J.A., 1982, *Gravitational lenses and cosmological evolution*. *Mon. Not. R. Astron. Soc.* 199, pp. 987–1006.
- PEEBLES, P.J.E., 1973, *Statistical analysis of catalogs of extragalactic objects. I. Theory*. *Astrophys. J.* 185, pp. 413–440.
- PEIRCE, C.S., 1887, *Criticism of "Phantasms of the Living": An examination of the Argument of Messrs. Gurney, Myers and Podmore*. *Proc. Am. Soc. Psych. Res.* 1 (1885–89), pp. 150–157. GURNEY, E., *Remarks on Professor Peirce's paper, Ibid.* pp. 157–179. PEIRCE, C.S., *Mr Peirce's rejoinder. Ibid.*, pp. 180–215.
- PENZIAS, A.A. and WILSON, R.W., 1965, *A measurement of excess antenna temperature at 4080 Mc/s*. *Astrophys. J.* 142, p. 419.
- REFSDAL, S., 1964, *The gravitational lens effect*. *Mon. Not. R. Astron. Soc.* 128, pp. 295–306.
- ROBSON, J., 1987, *Origin and Evolution of the Universe: Evidence for Design?* (McGill-Queens University Press, Montreal, Kingston).
- SCHNEIDER, P., 1986, *Statistical gravitational lensing and quasar-galaxy associations*. *Astrophys. J.* 300, pp. L31–L34.
- TURNER, E.L., 1980, *The effect of undetected gravitational lenses on statistical measures of quasar evolution*. *Astrophys. J.* 242, pp. L135–L139.
- VAUCOULEURS, G. DE, 1971, *The large scale distribution of galaxies and clusters of galaxies*. *Publ. Astron. Soc. Pacific* 83, pp. 113–143.
- WALSH, D. et al., 1979, 0957 + 561 A, B: *twin quasistellar objects or gravitational lens?* *Nature* 279, pp. 381–384.

- WEEDMAN, D.W. *et al.*, 1982, *Discovery of a third gravitational lens*. *Astrophys. J.* 255, pp. L5–L9.
- YOUNG *et al.*, 1980, *The double quasar Q09571 + 561 A, B: a gravitational lens image formed by a galaxy at $z = 0.39$* . *Astrophys. J.* 241, pp. 507–520.
- YU, J.T. and PEEBLES, P.J.E., 1969, *Superclusters of galaxies*. *Astrophys. J.* 158, pp. 103–113.
- ZWICKY, F., 1937, *On the probability of detecting nebulae which act as gravitational lenses*. *Phys. Rev.* 51, p. 679.

PROBABILITY IN DYNAMICAL SYSTEMS*

JAN VON PLATO

Department of Philosophy, University of Helsinki, Helsinki, Finland

Probability theory is applied mathematics. Finding new scientific applications for it has been a major factor in its development. It is difficult to do justice to these applications, if we adopt the view that the concept of probability always has one and the same meaning. Instead of a single interpretation, one can pose questions pertinent to a particular application. Then, the interpretation is not given *a priori*, but requires a justification. Answering the right questions, we should come to know what the basis of our interpretation of probability is in the application concerned. To some extent, interpretation becomes a question of scientific research into the particular case studied.

Probabilities are numbers between zero and one that are additive in a special way. The probability that one of two alternative events occurs can be computed as a sum of the probabilities of the alternatives, minus the probability that the alternatives occur simultaneously. This guarantees that probability numbers always remain within the above bounds. An interpretation of probability should give meaning to these numbers. It should tell us what they can be used for. It should tell us what the things, events or whatever, are, to which probability numbers are attached. Most importantly, an interpretation should tell us *how we arrive at these*

* This essay differs considerably from the invited talk at the Eighth International Congress of Logic, Methodology and Philosophy of Science, Moscow 1987. It was given in the present form during a stay at the Department of Philosophy of the University of Bologna in October and November 1987. I am grateful to Professor Maria Carla Galavotti for providing a fruitful working atmosphere that led to the present much improved conception. I also wish to take the opportunity of expressing my warmest thanks to Professor Abner Shimony whose perceptive remarks, in his (1975), first brought my attention to this field of inquiry, and to Professor Isaac Levi who for ten years by now has challenged my aprioristic persuasions in the philosophy of probability.

probability numbers. What is the particular way of determining probabilities in a given kind of application?¹ If you have an idea of how probabilities are determined, you can start checking whether the probability of some specific event is some specific number. Then, knowing what the grounds are for asserting that the probability of an event is such and such, you know what probability means in the situation considered. What difference does it make for events to have different probabilities, say one in a million or one half? Can different grounds for determining probabilities be compared in any way, say one being better informed than the other, or one being right and the other wrong? If a suggested interpretation of probability does not address questions of this kind, if it remains silent about the determination of probability numbers, it is difficult to say in what sense it gives meaning to probability at all. One sometimes hears probability is a normalized measure, so that probability numbers are values of normalized measure functions. This is precisely the kind of answer that leaves the determination of values of probabilities completely open. It is rather useless to anyone who has to apply probability theory. It is just saying in a very general way (say, for Banach spaces) that probabilities are those numbers between zero and one that are additive in a special way.

Until the middle of the last century, the application of probability in physics was based on the idea of a true—if unknown—actual course of events. Probabilities were thought of as expressing to *what degree* the true course of events is unknown. Particularly, this was the case for the theory of errors. There the use of probability was thought to reflect ignorance of the true value of a physical quantity. Measurement of the unique true value was marred by difficulties that were thought theoretically unimportant.

Probability made its entrance into physics in the 1850s. Herschel, in a review of Quetelet's work, suggested a way of deriving a two-dimensional normal distribution. It contained a conceptual novelty. He considered an iron ball that is repeatedly dropped on the floor. Assuming that the deviations from an absolutely vertical fall are independent in some orthogonal x - and y -directions, a normal distribution follows. What is new and particular about the situation is that the variation is real here. The event itself is of a repetitive character, not only the measurement.

¹The significance of new ways of determining probabilities, both historical and philosophical, is one of the leading themes of COSTANTINI (1980). The influence of this point of view should be clear in the following.

There is no unknown true value. The argument was taken up by Maxwell. He substituted molecules of a gas for the iron ball. This led to a Maxwellian distribution law from as one now says, spherical symmetry of the molecular motions.²

The application of probability theory in statistical mechanics received very early on two interpretations. In the words of Boltzmann himself: "There is a difference in the conceptions of Maxwell and Boltzmann in that the latter characterizes the probability of a state by the average time the system is in this state, whereas the former assumes an infinity of equal systems with all possible initial states."³ The use by Maxwell of what we now call ensembles has always been subject to interpretation and criticism. Some physicists, such as R.C. Tolman in his influential book of 1938, take their use as an expression of genuine *a priori* probabilities. Many people think this is the view behind Gibbs' virtual ensembles. But if you take up his book, you find him saying: "It is in fact customary in the discussion of probabilities to describe anything which is imperfectly known as something taken at random from a great number of things which are completely described. But if we prefer to avoid any reference to an ensemble of systems, we may observe that the probability that the phase of the system falls within certain limits at a certain time, is equal to the probability that at some other time the phase will fall within the limits formed by phases corresponding to the first."⁴ The probability is stationary here; it is a feature of the system invariant in time. This invariance, stationarity, is basic to the Boltzmannian interpretation of probability as a limit of time average. It is strange that the time average interpretation has received so little attention in the philosophy of probability. It was Einstein's notion of probability, and a very crucial one as it made fluctuations into physically real phenomena. These occur with the time development of a single system. They are not uninterpreted dispersion terms for a physically inexistent ensemble. Lorentz and with him a host of Dutch statistical mechanists shared Einstein's interpretation. Marian von Smoluchowski's essay of 1918 is one of the highlights of these developments. Following Poincaré, he gave a physical definition of chance. It has its origin in the instability of dynamical motion. Any slight change in initial conditions will lead to macroscopically different behaviour.⁵

² This matter is discussed in EVERITT (1974: pp. 136–137). The work of Clausius, another early champion of probabilistic arguments in physics, also is treated there.

³ BOLTZMANN (1909: p. 582).

⁴ GIBBS (1902: p. 17) of the Dover edition.

⁵ See VON PLATO (1987) and VON PLATO (1983a) for these developments.

The precise mathematical justification of what physicists were commonly doing, namely the equating of phase averages with time averages and the computation of the former so as to determine the latter, has not been a very successful enterprise. That this is so, and was so, did not concern Einstein and others who were convinced that plain ordinary gases have total energy as their sole permanent physical property (dynamical invariant). The first positive results on this *ergodic problem* came in the early thirties. By then, the classical theory had lost much of its physical importance. On the other hand, the study of probability in dynamical systems soon led to a general, purely probabilistic formulation of the theory. One attractive aspect of this theory is the following. Powerful concepts are formulated in it; and the ensuing results are usually interpretable in some suitable systems of classical dynamics. One is therefore able to justify the probabilistic assumptions of the general theory from the physical description of the particular classical case studied. In the general theory, these remain hypothetical.

1. Probability: interpretations

Interpretations of probability are broadly divided into epistemic and objective types. In the first type, probabilities are *degrees of belief*. One arrives at them through the introspection of one's state of mind, concerning something one is uncertain about. A particular way to measure degrees of belief is to force the person in question into a comparison: He is told to make a choice between betting for the occurrence of the event of interest, and bets in a standard system as found in, e.g., lotteries. By refining the latter system, one arrives at more and more detailed figures. The most common objectivist view, on the other hand, says that probabilities are (or are almost) the same as limits of relative frequencies. They are estimated from data according to a well established statistical methodology.

Neither of the above views needs to be concerned about chance; whether it exists and in what sense. The epistemic or subjectivist philosophy can remain non-committal as to the true grounds for the existence of uncertainty in people's minds. Also, the limiting frequencies do not alter an inch if they are limits of, in some sense, deterministic sequences. Historically, though, the epistemic interpretation has been connected with Laplacian determinism. The statistical interpretation often came with a requirement of randomness, as in the theory of von Mises.

Some of the problems connected with the above views are: the epistemic interpretations fail to account for the obvious reality of statistical laws in Nature. Frequentism limits probabilities to repetitive phenomena, and strictly taken only to those form which we can obtain data. A specific problem concerns the strong laws of large numbers and other limit theorems. They say a certain probability law holds for a continuous set of denumerable sequences of results, with probability one. The latter probability cannot be interpreted frequentistically, as getting just one elementary event from the continuous set of all possibilities takes up an infinity of time. We conclude that anyone who advocates a single and exclusive interpretation of probability for all cases, is in trouble. Either he is led into artificialities, or into denying the reasonableness of what many people think are legitimate uses of probability.⁶

I shall now add to these troubles by discussing one more interpretation of probability. I shall denote it by what one sometimes sees in the physical literature, namely:

2. Probability as time average

We shall need some concepts from the theory of dynamical systems. It will be enough to think always of the example of a finite dimensional real space R^n . There will be

- A *state space* S : this is the set of all physically possible states of a dynamical system. Typically $S \subset R^n$.
- Elements of S : these are the exact states or *microstates* x, y, z, \dots of the system.
- A *dynamical law*: if at time t_0 our system is in microstate x_0 , it will evolve according to its dynamical law of motion, for any given time t_1 , into some definite microstate x_1 . We write this basic dynamical relation as $x_1 = T(t_0, t_1, x_0)$.

Next assume:

- Given that the system is in initial state x_0 at time t_0 , the law T will always act in the same way, irrespective of what the initial time t_0 is.

⁶ DE FINETTI is the most consequent example of a subjectivist who totally abandons the idea of objective probability (cf. references in note 10 below). The frequentist VON MISES (1972: p. 88) thinks subjective probability is based on a "Denkfehler".

In this case we may write $T(t_0, t_1, x_0) \equiv T(t, x_0)$, where t is the length $t_1 - t_0$ of the time interval under which we follow the motion. The dynamical evolution of a state x forms a *trajectory* in state space, as given by $T(t, x)$ as a function of t . The dynamics is time independent. This can also be expressed by saying that the law of motion of the system remains the same in time. It obtains physically when a system has a constant total energy, that is, the system is energetically isolated. We also assume that the set of possible states S is of finite measure, typically it is such a connected subset of R^n for some n . A few examples are:

1. Plain ordinary gas in a container with $n \approx 10^{24}$.
2. Kepler motion of two gravitating point masses, with $n = 6$. The center of mass is held fixed in the origin, so that three position and three momentum (or velocity) coordinates are needed for one point.
3. Points moving in a unit square, with $n = 4 \times$ number of points.

Let us now assume that S is given by the unit square. Call the set of states in the quarter $0 \leq x, y \leq \frac{1}{2}$ the *macrostate* A . All the points that can be reached from state x form the trajectory of x . From some zero time 0 to time t , check the relative time the trajectory of x is in the quadrant under question. Call the latter the *occurrence* of macrostate A . The relative time is given by the integral

$$(1/t) \int_0^t I_A(T(t, x)) dt$$

where I_A is the indicator function of the set A . Next let time go to infinity. Then, supposing it exists,

$$\lim_{t \rightarrow \infty} (1/t) \int_0^t I_A(T(t, x)) dt = \hat{I}_A(x)$$

is the limit of time average of the function I_A along the trajectory of the microstate x . In classical dynamics, it can be shown that the limit exists exactly when the system is isolated and has a time independent dynamics. Systems of this kind are called stationary.⁷

The limit of time average $\hat{I}_A(x)$ depends on (varies with) the trajectory.

⁷ Concepts and results referred to can be found in, e.g. FARQUHAR (1964) which is physically oriented. A recent extensive book is CORNFELD *et al.* (1982). Ergodic theory is treated probabilistically in BILLINGSLEY (1965). Newer results can be found in ORNSTEIN (1974).

However, there is a condition which guarantees that it is in fact a constant over S . Call a set B *invariant* if no trajectories lead in or out of B : let $T(t, B)$ be the image of the set B under the transformation $T(t, x)$. Invariance is now defined as $T(t, B) = B$ for all t . If one assumes that S and the empty set are the only invariant sets of a stationary system, the system is *ergodic*. Any set (of positive measure) is visited by all (almost all in the sense of measure theory) trajectories of the system. Individual trajectories winding in this complicated way are also called ergodic. The following result, called the ergodic theorem, shows the importance of the notion of ergodicity. For ergodic systems, we have

$$\hat{I}_A(x) = \hat{I}_A = \text{constant.}$$

The limit of time average of a macrostate A is a constant; it is *independent* of the particular dynamical evolution taking place.

How does probability enter the above picture? Let us assume that the point of our unit square is a coordinate pair for the position and momentum of a point mass in a Hamiltonian formulation of mechanics. There is a measure, called the natural measure or sometimes the microcanonical measure, over the Hamiltonian state space. This latter is usually called a phase space. The measure is defined by the differential element dx of our phase space. Then, assuming our system has a law of motion such that the property of ergodicity is fulfilled, it follows that the time averages of macrostates coincide with their microcanonical measures. Let P be the measure, normalized so that $P(S) = 1$. Then

$$\bar{I}_A = \int_S I_A(x) dx = P(A).$$

The numbers $P(A)$ have the properties of probability. For our above case, if it were shown ergodic, we could compute

$$\bar{I}_A = P(A) = \frac{1}{4}.$$

The original goal of the ergodic theory of classical dynamical systems was a determination of their statistical laws, starting from a dynamical description. It has been often said that one cannot get probabilistic conclusions from mechanical premises. But here this almost seems to succeed. The integral of the phase function I_A over S is expressed entirely in dynamical terms. The statistical element of the derivation is in the assumption that S is the right set of *possible* microstates. The measure

over S has to respect this notion of possibility or physical accessibility. One way of putting the matter is: If you fix which sets have measure zero, the dynamics will sort out the right measure function P under which the transformations $T(t, x)$ preserve this measure. In this sense, all other probability numbers $P(A)$ except those which are zero or one, are computed from the dynamics.

Now we know how time average probabilities are determined. We should be able to answer to some of the kinds of questions about interpretations of probability discussed above:

- The probability numbers attach to macrostates of suitable dynamical systems. They are good for predicting the asymptotic (if not shorter) time averages of such systems.
- We arrive at these numbers as follows. First, they only exist if the system is stationary which is a physical condition of obvious meaning. Secondly, they are unique if the system is ergodic which again is a physical condition. Finally, if one has the right set of physically possible states, the probability measure is determined from the dynamical equation.

The question of chance can also be addressed. It is strictly speaking physically meaningless to require that one waits an infinite time to see if the predictions of the theory were correct or not. The statistical laws one derives are only related to the asymptotic behaviour of the system. For finite times, one needs stronger properties than ergodicity. Then, chance is defined as instability of the dynamical behaviour of a system. There are systems whose behaviour in time is so strongly non-linear that their macroscopic behaviour is that of a random process with probabilistic independence of consecutive macrostates. This kind of wiping out of any memory of a system's past in its behaviour in time comes from the exponential rate of separation of initially close trajectories. Then, prediction of how the system exactly behaves becomes impossible in principle. It would require an infinitely precise determination of the initial conditions of the system, which is physically meaningless.

3. Comparison with frequentist probability

Time averages obviously are some kind of continuous counterparts to limits of relative frequencies. By discretizing the phase space and by

considering the occurrence of events at discrete intervals of time, one gets directly limits of relative frequencies from limits of time averages. However, the former are a very special form of limiting frequencies as we shall see.

Let us divide our unit square by the line $x = \frac{1}{2}$. The left side is denoted as macrostate B . Let t be a chosen unit of time. Write $T(t, x) = T(x)$, $T(2t, x) = T^2(x)$, $T(nt, x) = T^n(x)$ and so on. We get a discrete form of trajectory $(x, T(x), T^2(x), \dots)$ from the continuous one given for the state x by $T(t, x)$. The limit of time average of the macrostate or event B becomes a sum:

$$\hat{I}_B(x) = \lim_{n \rightarrow \infty} (1/n) \sum_{i=0}^{i=n-1} I_B(T^i(x)).$$

This number is the limit of a relative frequency. But it is of a special kind. The sample sequence, as one says in statistics, is a dynamically determined process. The questions of the existence and uniqueness of limits of relative frequencies are well defined and can be in principle settled on the basis of properties of the dynamical law.

We conclude that the discretization of a stationary or ergodic dynamical system does not merely lead into limits of relative frequencies. For a first thing, here the sampling is performed by the dynamics. Secondly, to put some of our remarks in a statistical jargon, the requirement of unbiased sampling is met. Ergodicity is at least asymptotically the proper guarantee for having representative sampling. For the ergodic theorem proves precisely that the statistical properties are invariant features of a system, and independent of the particular trajectory.

A third and most salient difference is the following. In statistics, frequentist probabilities are determined through estimation from data, or from a probabilistic model that has been tested against observations and found to be satisfactory. In the case of limiting frequencies obtained by discretizing time averages, it is *the other way around*. One seeks to compute the limiting frequencies from the physical description of the system under study. No data are needed here. Therefore, we emphasize again the important role played by ways of determining probability numbers, in questions of interpretation of the concept of probability. That these theoretically determined numbers coincide with limits of time averages or relative frequencies, is only a consequence of the theory's being right about the time behaviour studied. The numerical coincidence does not subsume probabilities in dynamical systems under the notion of frequentist probability.

The time average probabilities of dynamical systems differ in at least the above three ways from frequentist probabilities. While keeping this in mind, it is still useful to compare the notions from statistics in general with the well-defined and precise concepts of the ergodic theory of dynamical systems. The following tables provide such a comparison.

First we have notions connected with the basic description of the situation:

STATISTICS:	DYNAMICS:
Sample space	Phase space
Elementary event	Microstate
Event	Macrostate
Random variable	Phase function

Phase functions include all the physical properties of a system.

Secondly we have notions connected with sampling or time behaviour:

Sampling process	Dynamical law
Finite sample	Finite trajectory
Frequency	Time average
Limit of frequency	Limit of time average

As a third group we have notions connected with structural properties:

“Same circumstances”	Stationary system
Fair sampling (asymptotic condition)	Ergodic system
Independent repetition	Bernoulli system

Bernoulli systems are the last ones in a list of randomness properties of dynamical systems. The comparison could be continued by setting the notion of a sufficient statistic in the left column, and the notion of a complete system of invariants in the right. However, of these concepts only the latter one will be discussed in this essay. We shall treat in order some problems and results in the foundations of probability theory for which the dynamical framework seems especially relevant. Characteristically, as was noted above, one can always work out examples where the

concepts and results receive a physical interpretation. The applications below are about the notion of randomization, and about the description of experimentation from a dynamical point of view.⁸

4. Randomization and mixtures⁹

The ergodic theorem is not only a sufficient condition for the uniqueness of limits of time averages. It is also a necessary condition. The general probabilistic formulation of ergodicity, the non-existence of non-trivial subsets invariant under a transformation representing the repetition of experiments, is consequently a characteristic of situations in which a frequentist probability can be introduced.

Let us consider the interpretation of the case where we have only a stationary, non-ergodic measure P over a dynamical system. Time averages such as the $\hat{I}_A(x)$ above exist, but vary in general with x . If we compute the phase average of $\hat{I}_A(x)$ over the state space S , we get

$$\bar{\hat{I}}_A(x) = \int_S \hat{I}_A(x) dP = P(A).$$

If we had the possibility of sampling states x according to the probabilistic law P , and if we followed the time averages $\hat{I}_A(x)$, the average of these time averages would reproduce what could be called the *a priori* probability $P(A)$ of A .

There is a beautiful result in ergodic theory which says that if you have a stationary system, its probability law P can be represented as a mixture or integral over the laws for the parts of the system that have the property of ergodicity. This representation is unique. The state space of a stationary system decomposes, as one says, into ergodic components. Each of these carries its own ergodic measure, and the probability law of the whole is a weighted average of them. One can think of the components as statistically homogeneous subpopulations of a population in statistical equilibrium as a whole. Let us say we have only a finite or denumerable set of ergodic components, that is, invariant sets which do not have any non-trivial invariant subsets. Then there is a unique probability law for each component, indexed by a parameter λ_j . This parameter itself is

⁸ See VON PLATO (1988) for some further applications.

⁹ I borrow the title of this section from FELLER (1971: §II.5). His suggestive remarks provided much of the inspiration for the present considerations.

subject to a uniquely defined probability distribution μ . If our system is prepared in a way which corresponds to the i th component, it is described statistically by the law P_{λ_i} . If we wish to sample from the whole population, we must *randomize* with respect to the factorizing parameter λ . Then, the probability of an event A is given as a mixture. With a finite or denumerable number of values for λ , we have

$$P(A) = \sum_i a_i P_{\lambda_i}(A)$$

where the weights a_i are uniquely determined from P . The a_i are the probabilities for the values of λ . In the continuous case it is as in the previous integral expression, only with a mixture over the range of λ instead of S . Physically speaking, we must repeatedly disturb the system somewhat, or prepare it repeatedly, to make it jump from one invariant set into another, and do this according to the probability law μ .

Matters of the above kind have as one of their special cases a result that has received very much attention in the foundations of probability. I refer to what is known as de Finetti's representation theorem. It says exchangeable measures are unique mixtures of independent measures.¹⁰ De Finetti thinks there is no true randomization. There is only a (as he says, fictive) true unknown parameter value. For simple experiments, it coincides with the (fictive) true unknown objective probability of an event. Consecutive events are not probabilistically independent under exchangeability. By conditionalization, one identifies at least asymptotically the probability law of the component one is in. What the statistical laws P_λ of the other components are, remains unknown, or may be even meaningless as these other values do not become operative in any way.

The situation is described differently if we randomize. Then, the probability of a single result is again mathematically given as a mixture of the form shown above. But now it should be seen as a conditional probability. One first chooses at random (sic) a value for λ . Then one performs the experiment, trial or whatever. At each repetition, *both* operations are repeated. In consequence, the repetitive structure of

¹⁰ The result and its philosophy are by now so standard that it should suffice with a few references for those uninitiated. De Finetti's own philosophy is in his classic *La Prévision* (1937), English translation in DE FINETTI (1964). My own views can be found in VON PLATO (1983b). For the by now extensive field of exchangeability mathematics, see the recent review of ALDOUS (1985).

sequences of results is described probabilistically by a product measure, i.e., the successive results are independent.

Now we see that stationary, non-ergodic probabilities can be interpreted in different ways. A special case is where you have exchangeability and independence. The former is a special case of stationarity, the latter a special case of ergodicity. You can argue that only one component is real, and the others fictive. Then, you might think the mixing measure μ represents your degree of ignorance as to the true probability law P_λ . Note that this is *not* de Finetti's position, although it comes close to that. For him, nothing but the subjective degrees of belief are real. A second situation is where the randomization is real. Or even, we have cases where the different components physically exist simultaneously, and then the mixture represents the probabilistic law of the whole population. An example is the phase decomposition of matter.¹¹ There, the weights a_i come from the relative volumes of the different phases of matter.

5. A dynamical approach to the description of experiments

In a Hamiltonian formulation of mechanics, the order in which one conceives of the state space is a reverse of the traditional. There, one thinks the space consists of its points, and the set of trajectories of the individual trajectories. In a Hamiltonian formulation, we start with as one says a global picture of the space as a whole.¹² Parts of it are identified by *invariant phase functions*: these are the functions f such that $f(T(t, x)) = f(x)$. For stationary systems, the total energy $H(x)$ is an invariant, since it is constant. If the dimension of the phase space is n , the value $H(x) = H$ constrains the trajectories to lie on a hypersurface of constant energy, so one dimension is reduced. Proceeding further, one constrains the motion into subspaces of lower and lower dimensions. In our examples above, we had $n \approx 10^{24}$ for plain ordinary gas. As there are no further invariants, there are no physical means for constraining the motion to remain in anything else but the whole hypersurface of constant energy. In the second example of Kepler motion, there are invariants that constrain the motion into orbiting an ellipse in configuration space. This leaves only one degree of freedom, so the ergodic components are one-dimensional

¹¹ See VON PLATO (1983b) and references to physics literature therein.

¹² See BALESCU (1975) for a review of Hamiltonian dynamics (chapter 1) and of the ergodic problem (Appendix).

sets, individual trajectories. Here, the traditional picture of “deterministic motion” works physically.¹³

Phase functions represent properties of dynamical systems. The constant total energy is the most important of these. It is an invariant property, one that remains the same in time. Other permanent properties are of course also represented by invariant functions. Any other properties a system might have, are contingent and wiped out by the dynamical evolution of the system. For stationary systems, the values of time averages do not depend on when one starts averaging. Therefore time averages are invariants. That is, probabilities are permanent physical properties in these cases.

Invariant functions correspond uniquely to invariant sets. The indicator function I_A of an invariant set (this required, $T(t, A) = A$) is an invariant function. If a function f on the other hand is invariant, the set $f^{-1}[x] = \{y | f(x) = f(y)\}$ is invariant. A system of invariant functions f_1, \dots, f_k is a *complete set of invariants* if all other invariants can be expressed as functions of these. For our two examples above, we have $k = n - 1$ for Kepler motion. This is the “deterministic” case. For the case of plain ordinary gas we have $k = 1$.

In both cases we have a family of probability measures P defined over the phase space. It is parametrized by values $\lambda = (\lambda_1, \dots, \lambda_k)$ of the invariants f_1, \dots, f_k . Then each set of values gives an invariant, ergodic component, which bears the probability measure P_λ . For the system of Kepler motion, the measure is concentrated on a single trajectory. For the other cases it is usually defined over a continuous set.

What a complete set of invariants does is obviously the ergodic decomposition of stationary probability measures discussed above. In fact, this is how von Neumann came to the ergodic decomposition theorem some 55 years ago. Its general probabilistic significance became clear somewhat later, when Hopf and Khintchine realized that the ergodic theorems can be given independently of the classical dynamics.

¹³ As was said, the traditional order is to think of the state space as being formed of its points. These are thought to pre-exist individually, irrespective of any constructive means of identifying them. Then, bad (classical) logic leads one into postulating for arbitrary subsets A of S and for any state x the principle $x \in A$ or $x \notin A$. Putting the points of a trajectory (existing in a set theoretic sense) in place of A leads to traditional determinism. It follows by sheer classical logic which says things are this or that way irrespective of what we ever are able to know. The Hamiltonian approach on the other hand is based on physical identifiability, not set theoretical existence, of parts of phase space. The example of Kepler motion is discussed in STERNBERG (1969).

Our remarks on randomization and mixtures could have been equally well given in terms of invariants.¹⁴

If we consider the discrete case, invariants work in the same way. The sample sequences have an underlying dynamical trajectory so that the invariants are determined as soon as their dynamics is given. Invariants are sometimes called controllable integrals in the study of dynamics. In the statistical and philosophical literature, on the other hand, one speaks of control parameters and causal factors. This is exactly what can be done with invariants. They can be used for controlling the permanent properties of time evolutions of systems. These can be statistical laws, as when you have for the number k of independent invariants the inequality $k < n - 1$. Or there may obtain "deterministic causality" where $k = n - 1$. Then, the values of the factors being fixed, the determination of one contingent property is enough for the identification of a unique evolution. In the statistical case, what is caused by the choice of factor values is statistical behaviour, so this situation could be called that of probabilistic causality. In no way are the two cases with either $k < n - 1$ or $k = n - 1$ opposites of each other; the latter form of causality is one where the parametric probability measure is concentrated on a single trajectory. The ideal of research is, in terms of our scheme of dynamical systems, a situation where all the invariants have been positively identified. Then, in the statistical case, it is possible to determine the probabilistic laws of a component or subpopulation, knowing at the same time that there exist no further factors which could be used for altering the statistical behaviour obeying those laws.

The contrary of the above, the case of what could be characterized as *incomplete information*, has also a place in our philosophy of probability in dynamical systems. It obtains in a case where we are *unable* to identify all the invariants we otherwise know must exist. Then we positively know there must be incomplete information. This we represent by introducing a subjectively interpreted *a priori* probability at the level (subspace) we have been able to identify. Assuming $k < n - 1$, the behaviour of the system still belongs to the statistical type. It follows that the statistical behaviour depends on factors whose values remain accidental from the experimenter's point of view.

¹⁴Systems with an arbitrary number of invariants are first treated systematically in GRAD's memoir (1952). A measure theoretic treatment is given in LEWIS (1960). TRUESDELL (1961) gives a popular exposition. These works do not make explicit use of the ergodic decomposition point of view, which we find somewhat surprising.

The traditional idea was different from the above. According to the traditional view, there is incomplete information as long as we have anything less than single trajectories. Therefore this view considered probabilities in classical systems always as expressions of ignorance of the true deterministic course of events. Now we know better: beyond a complete set of invariants, *there is no information* to be missed.

References

- ALDOUS, D.J., 1985, *Exchangeability and related topics*, Lecture Notes in Mathematics, Vol. 1117, pp. 1–198.
- BALESCU, R., 1975, *Equilibrium and Non-Equilibrium Statistical Mechanics* (Wiley, New York).
- BILLINGSLEY, P., 1965, *Ergodic Theory and Information* (Wiley, New York).
- BOLTZMANN, L., 1909, in: F. Hasenöhl, ed., *Wissenschaftliche Abhandlungen*, Vol. II (Leipzig).
- CORNFIELD, I.P., FOMIN, S.V. and SINAI, YA. G., 1982, *Ergodic Theory* (Springer, Berlin).
- COSTANTINI, D., 1980, *Inductive probability and inductive statistics*, in: M.L. Dalla Chiara, ed., *Italian Studies in the Philosophy of Science* (Reidel, Dordrecht).
- EVERITT, C.W.F., 1974, *James Clerk Maxwell Physicist and Natural Philosopher* (Scribner's, New York).
- FARQUHAR, I.E., 1964, *Ergodic Theory in Statistical Mechanics* (Wiley – Interscience, New York).
- FELLER, W., 1971, *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd edn. (Wiley, New York).
- DE FINETTI, B., 1937, *La prévision: ses lois logiques, ses sources subjectives*, *Annales de l'Institut Henri Poincaré* 7, pp. 1–68.
- DE FINETTI, B., 1964, *Foresight: its logical laws, its subjective sources*, in: H.E. Kyburg and H. Smokler, eds., *Studies in Subjective Probability* (Wiley, New York), pp. 97–158.
- GIBBS, J.W., 1902, *Elementary Principles in Statistical Mechanics* (as republished by Dover, New York 1960).
- GRAD, H., 1952, *Statistical mechanics, thermodynamics and fluid dynamics of systems with an arbitrary number of integrals*, *Communications in Pure and Applied Mathematics* 5, pp. 455–494.
- LEWIS, R.M., 1960, *Measure-theoretic foundations of statistical mechanics*, *Archive for Rational Mechanics and Analysis* 5, pp. 355–381.
- VON MISES, R., 1972, *Wahrscheinlichkeit, Statistik und Wahrheit*, 4th edn. (Springer, Vienna).
- VON NEUMANN, J., 1932, *Zur Operatorenmethode in der klassischen Mechanik*, *Annals of Mathematics* 33, pp. 587–642, with additions, *ibid.*, pp. 789–791.
- ORNSTEIN, D.S., 1974, *Ergodic Theory, Randomness and Dynamical Systems* (Yale University Press, New Haven).
- VON PLATO, J., 1983a, *The method of arbitrary functions*, *The British Journal for the Philosophy of Science* 34, pp. 37–47.
- VON PLATO, J., 1983b, *The significance of the ergodic decomposition of stationary measures for the interpretation of probability*, *Synthese* 53, pp. 419–432.
- VON PLATO, J., 1987, *Probabilistic physics the classical way*, in: L. Krüger, G. Gigerenzer

- and M. Morgan, eds., *The Probabilistic Revolution*, Vol. 2 (MIT Press, Cambridge), pp. 379–407.
- VON PLATO, J., 1988, *Ergodic theory and the foundations of probability*, in: B. Skyrms and W. Harper, eds., *Causality, Chance, and Credence*, Vol. 1 (Kluwer, Dordrecht), pp. 257–277.
- SHIMONY, A., 1975, *Carnap on entropy*, in: J. Hintikka, ed., *Rudolf Carnap, Logical Empiricist* (Reidel, Dordrecht), pp. 381–395.
- VON SMOLUCHOWSKI, M., 1918, *Über den Begriff des Zufalls und den Ursprung der Wahrscheinlichkeitsgesetze in der Physik*, *Die Naturwissenschaften* 6, pp. 253–263.
- STERNBERG, S., 1969, *Celestial Mechanics*, Part I (Benjamin, New York).
- TOLMAN, R.C., 1938, *The Principles of Statistical Mechanics* (Oxford).
- TRUESDELL, C., 1961, *Ergodic theory in classical statistical mechanics*, in: P. Caldirola, ed., *Ergodic Theories* (Academic Press, New York), pp. 21–56.

This Page Intentionally Left Blank

8
Foundations of
Physical Sciences

This Page Intentionally Left Blank

AN AXIOMATIC BASIS AS A DESIRED FORM OF A PHYSICAL THEORY

G. LUDWIG

Department of Physics, University of Marburg, D-3550 Marburg, FRG

Let us regard a physical theory PT as given in the following form: PT is composed of a mathematical theory MT , correspondence rules (\dashv) and a reality domain W .

The correspondence rules are prescriptions of how to translate into MT those facts which can be detected in nature or on devices, or arise by technical procedures. Only facts in the fundamental domain G (a part of W) shall be considered. It is important that the description of facts in G does not use the theory under consideration. This does not mean that we use no theory at all. But we may use only “pretheories” to describe the fundamental domain G , that is, theories already established before interpreting the theory PT to be considered.

Clearly the fundamental domain G must be restricted to those facts which may be translated into the language of MT . However, further restrictions of G are often necessary. Such restrictions can be formulated as normative axioms in MT . All those facts must be eliminated from the fundamental domain which, when translated by the correspondence rules, contradict the normative axioms.

The correspondence rules have the following form: some facts in G are denoted by signs, say letters a_1, a_2, \dots . In MT some sets are singled out as pictorial sets E_1, E_2, \dots, E_r and some relations R_1, R_2, \dots as pictorial relations.

By virtue of the correspondence rules the facts of G are translated into an additional text in MT (sometimes called “observational report”), which is of the form

$$\begin{aligned} & (\dashv)_r(1): a_1 \in E_{i_1}, a_2 \in E_{i_2}, \dots \\ & (\dashv)_r(2): R_{\mu_1}(a_{i_1}, a_{k_1}, \dots, \alpha_{\mu}), R_{\mu_2}(\dots) \dots, \\ & \quad \text{not } R_{\nu_1}(\dots), \dots \end{aligned}$$

Here the α_μ are real numbers which may be absent in some of the relations R_μ . Let A denote the total text $(-), (1) (2)$.

The distinction between $(-), (1)$ and $(-), (2)$ has no fundamental significance. Instead of $(-), (1)$ one may introduce relations $T_i(x)$ equivalent to $x \in E_i$.

The correspondence rules are just rules for the translation of propositions from common language or from the language of pretheories into the relations $(-), (1) (2)$. How this translation from pretheories should be done is shown by LUDWIG (1985–1987: XIII, §4.3).

MTA will denote the mathematical theory MT completed by the observational report A .

We say that the theory PT does not contradict experience if no contradiction occurs in MTA . We do not claim that this “no contradiction with experience” is the one and only criterion to accept a PT .

We know that in general the relations $(-), (1) (2)$ are not suitable to describe the facts in G because of imprecision. For the sake of simplicity, we will not give here the necessary generalizations of $(-), (1) (2)$ (see LUDWIG 1985–1987: XIII, §1).

I. Axiomatic bases

We have seen how the correspondence rules enable us to compare experiences with the mathematical theory MT , i.e. to establish MTA . Now we will write MT' instead of MT , A' instead of A and PT' instead of PT .

There arise several questions:

(1) How is the fundamental domain delimited? It could be that a contradiction in $MT'A'$ is only caused by the circumstance that in A' we had noted facts not belonging to the fundamental domain G .

2. Which part of MT' can in principle be rejected by experience? Which part of MT' are purely mathematical ingredients without any significance to experiments? For instance, do the axioms of MT' have any physical meaning?

The questions under (2) are crucial if MT' only consists of set theory (including the theory of real and complex numbers) and the pictorial sets E_ν , as the pictorial relations R_μ are constructed by using only the set of real numbers. Then all physical aspects of the theory must lie in the definitions of the pictorial sets E_ν and pictorial relations R_μ . How can we exhibit the physical aspects of these definitions?

We try to answer all these questions by introducing an axiomatic basis for PT' .

The starting point for an axiomatic basis is the singling out of the pictorial sets E_ν and pictorial relations R_μ . We construct a new mathematical theory MT , introducing basic terms y_ν (instead of the E_ν) and basic relations t_μ (instead of the R_μ) and an axiom $P(y_1, \dots; t_1, \dots)$, such that $P(E_1, \dots; R_1, \dots)$ is a theorem in MT' .

We construct a new PT from PT' as follows.

As fundamental domain G of PT we use G' of PT' . As mathematical theory we use MT . The correspondence rules are translated in obvious ways from PT' :

Instead of $(-)'_r(1)$ we write

$$a_1 \in y_{i_1}, \dots$$

Instead of $(-)'_r(2)$ we write

$$t_{\mu_1}(a_{i_1}, \dots), \dots$$

In this way, PT becomes a well-defined physical theory. It can be proved that " $MT'A'$ without contradiction" implies that MTA is also without contradiction.

A PT may be given (without recourse to PT' !) in the form that the base terms y_ν of MT are the pictorial terms and the basic relations t_μ the pictorial relations so that A is given by the above relations. Then we say that MT is an axiomatic basis of PT . Sometimes we shortly call PT an axiomatic basis.

The theory PT constructed above is an axiomatic basis. PT' can be stronger than PT , since we have only presumed that $P(E_1, \dots, R_1 \dots)$ (P the axiom of MT) is a theorem in MT' .

We say that the E_1, \dots, R_1, \dots represent the species of structure given by the terms $y_1 \dots; t_1, \dots$ and the axiom $P(y_1, \dots; t_1 \dots)$ if there is an isomorphism $y_i \xrightarrow{f_i} E_i$ where the t_μ are mapped onto the R_μ . If we have such a representation of the y_ν, t_μ by the E_ν, R_μ , the two theories PT' and PT are equivalent in the sense that MTA is without contradiction if and only if $MT'A'$ is without contradiction. Then we call also PT an axiomatic basis of PT' .

A very well-known example of a representation is that analytical geometry as MT' is a representation of Euclidean geometry as MT .

II. Laws of nature and theoretical terms

We have explained what is an axiomatic basis PT of a PT' . But we have not answered the question whether there is an axiomatic basis for every PT' .

The problem of finding an axiomatic basis for a given PT' can be solved in any case trivially. We only need to choose as axiom $P(y_1, \dots; t_1 \dots)$ of MT the relation, that "there are" terms (as the basic terms of MT') and isomorphic mappings $y_\nu \rightarrow E_\nu$ which map the t_μ onto the R_μ .

We feel that such an axiom is not yet in the form of a physical law that physicists have in mind. MT with such an axiom gives no new physical insight compared with MT' . Only the logical structure of PT with MT is simpler than that of PT' with MT' .

The usual form of quantum mechanics PT' takes a Hilbert-space H as a basic term of MT' . If we choose in MT the relation "there is a Hilbert-space H so that . . ." we get no new insight into the physical significance of the Hilbert-space-structure.

Those terms in the axiom P which are connected with an existential quantifier (that is: there is an x with $A(x)$) are called theoretical auxiliary terms. In this sense the Hilbert-space is a theoretical auxiliary term in quantum mechanics. Only if MT is an axiomatic basis we can define what is a theoretical auxiliary term. The same is the case for the following definitions.

Let us denote by "theoretical (not auxiliary) terms" all intrinsic terms in the axiomatic basis MT . The physical meaning of such terms is exactly defined by their deduction in MT and by the previously established physical meaning of the base terms y_1, \dots, y_r (that are the pictorial terms) and the basic relations t_μ (that are the pictorial relations).

The concept of theoretical term appears to us still too broad since the set of real numbers can enter arbitrarily. Let us now restrict the concept of theoretical term, allowing the real numbers to enter the terms only in the manner in which their physical meaning is given by the pictorial relations. Therefore we define: a physical term B is an intrinsic term in MT for which B can be constructed of the sets $y_1 \dots$ and relations t_1, \dots only. The physical interpretation of such a term B is given by the logical construction from the pictorial sets and pictorial relations and by additional conditions which can be formulated intrinsically.

Let us denote an axiomatic relation P in an axiomatic basis MT as physically interpretable, if in P there are only such quantifiers "there is $z \dots$ " for which z is a physical term.

How such a physically interpretable relation P can indeed be interpreted is expressed by the interpretation of the physical terms z and by the interpretation of the quantifiers “there is”. The last interpretation is not trivial and will be given later.

If we have an axiomatic basis MT and if P is physically interpretable we call MT a simple axiomatic basis. Such a simple axiomatic basis is exactly the desired form of a physical theory.

There are philosophers of science who believe that there are physical theories (e.g. the quantum mechanics of atoms) for which there is no simple axiomatic basis, i.e. for which theoretical auxiliary terms are inevitable. We have indeed given such a simple axiomatic basis for quantum mechanics in LUDWIG (1985–1987). In this axiomatic basis there are no such theoretical auxiliary terms as electrons, atoms, Hilbert-space and so on. All such terms are defined later as physical terms.

Having claimed a simple axiomatic basis as the “desired” form of a physical theory, we do not mean that other forms have no significance. On the contrary, we would, for instance, renounce analytic geometry and use only the Euclidean axioms.

We have additional requirements on the form of the axiom P . The form should make it possible to distinguish between those parts of P which are “pictures of real structures of the world” and those which are “descriptions of special concepts” or “prescriptions for actions”. But we will not discuss this here (see LUDWIG 1985–1987: XIII, §2.4 and 2.5).

By means of an axiomatic basis we can solve many problems as e.g. intertheory relations (see LUDWIG 1985–1987: XIII, §3). We will only discuss one problem: the concept of physically real and physically possible (see also LUDWIG 1985–1987: XIII, §4).

III. Physically possible and physically real

We use words as “real” and “possible” in many places in physics. But we use these words only intuitively without defining rigorously what we mean. Obviously the interpretation of a PT is not exhausted by the correspondence rules. These rules are only the foundation on which we can build a more comprehensive interpretation language.

Certainly we adopt the observational report as statement of real facts formulated in the language of MT . But we all know that we claim much more as real than is written down in the observational report. We use it to infer other, not observed realities. For instance, we speak of real electrons although only interactions of macrosystems are observed. We

speaking of real atoms which compose macrosystems. We speak of real electrons in a semiconductor. Is all this correct? Or are electrons and atoms in macrosystems only fictions suitable for explaining some properties of the macrosystems? All this can be clarified only if we have a rigorous method of proceeding from the observational report to other realities.

Much more mystical than the approach to realities seems to be our talk about possible facts though there is no sign in *MT* for such a logical category as “possible”, i.e. there is no modal logic in *MT*. Obviously “possible” is a word of the interpreting language, and it is one of the most important words. In physics we mostly do not ask “what is?” but rather “what is possible?”. This gives rise to the fundamental—but not physical—question: “What of everything possible should we realize?”

The intuitive usage of the words “real” and “possible” sometimes has led to errors. We need only mention questions such as: has a single microsystem a real state? Have the microsystems real properties? Can hidden variables be real? Is there something like a real propensity for every possible process, if we describe processes by probabilities?

These and many other questions make it necessary to develop a rigorous method for introducing such words as real and possible. The starting point is the introduction of hypotheses. Our concept of a hypothesis is more comprehensive than the usual one. We regard as given an axiomatic basis *MT* with y_1, \dots as base sets and pictorial sets and t_1, \dots as the basic relations and pictorial relations. *A* may be an observational report and *MTA* the theory with the observational report as additional axioms.

It is now possible to invent additional relations of the same form $x_1 \in y_{i_1}, \dots; t_{\mu_1}(x_{k_1}, \dots a_{k_1} \dots)$ as the observational report. For such additional relations we write *H*. *H* is called a hypothesis of the first kind. *MTAH* is the theory *MTA* with the additional axioms *H*.

We will extend this concept to more general hypotheses. Instead of $x_1 \in y_{i_1}$ we admit $x_1 \in T_{i_1}(y_1 \dots)$ where $T_k(\dots)$ is an echelon set, i.e. is constructed by product and power sets from the base sets. Instead of the t_μ we allow other intrinsically constructed relations for the x_i . Such an extended form of a hypothesis is called a hypothesis of the second kind.

Our definition of a hypothesis must be distinguished from a forecast. A hypothesis of the first kind can be related to imagined facts in the past and in the future. A hypothesis is not restricted to the future.

It is not possible to give here a complete classification of hypotheses. We will give only some concepts related to such a classification.

If *MTAH* is contradictory we call *H* “false”, otherwise “allowed”. “*H*

allowed" is equivalent to the statement that H can be added to MTA without contradiction. If not only H can be added without contradiction but also the relation "there are x_i with H " is a theorem in MTA we call H "theoretically existent". If in addition the family of the x_1, \dots is uniquely determined, we call H "theoretically existent and determined".

We will discuss here only the three cases: H allowed, H theoretically existent and H theoretically existent and determined.

If we have two hypotheses H_1 and H_2 we can define the composition H of these two hypotheses by H equal to " H_1 and H_2 " whereby we have to observe that we have to take distinct letters x in the two hypotheses.

We define H_1, H_2 as "compatible" if H_1, H_2 and the composite hypothesis H are at least allowed. If H_1 and H_2 are theoretically existent they are also compatible. There can be two allowed hypotheses of the first kind which are not compatible, a case very essential for physics.

The procedures just described concerning observational reports and hypotheses are beyond the scope of usual mathematics since such concepts as the field of hypotheses, allowed, and theoretically existent do not have definitions in the scope of a mathematical theory. For example, the field of hypotheses is not a set in a mathematical theory. The hypotheses are not existing things, they are made, made by us humans by applying the physical theory.

In this sense the mathematical framework of a physical theory is not a closed mathematical theory. Rather it is an open mathematical field within which we continually change the mathematical theories by observational reports and hypotheses. Only one part of all these theories is left unchanged: the axiomatic basis MT .

All this manipulation of mathematical theories within a PT shall be called the "mathematical game" of PT . Above we have given some rules of this game.

A physical theory does not only describe by the axioms of MT the physical laws which are "valid for ever". It also contains a variable part. Some aspects of this variable part are given by the mathematical game. And the development of this game depends essentially on our actions in playing this game.

Although the axioms of MT are not changed in the playing of this game, many of these axioms are already adjusted to the game.

For instance, they determine whether a hypothesis is theoretically existent or not.

This mathematical game is not played for the sake of itself. The game is highly significant for physics.

At the beginning of this section we said that it may seem mystical how

we can speak of possible facts though there are no corresponding logical signs in MT . This is different in the mathematical game. An allowed hypothesis may also be called “possible”, i.e. permitted to be added without contradiction. But we are not interested in possible moves in the purely mathematical game. We are interested in physics, i.e. in a physical interpretation of this game.

Just as a physical interpretation of a mathematical theory MT is not given by MT itself, also the physical interpretation of the mathematical game is not given by this game. The physical interpretation of MT was given by the correspondence rules, which permit us to write down the observational report in the language of MT . We now wish to extend this interpretation to the mathematical game, using the classification of hypotheses from the preceding sections.

Let us start by explaining what we mean by a “comparison of a hypothesis with experience”.

Here we begin with the simplest but also fundamental case, that of an allowed hypothesis of the *first* kind.

The simplest case for a comparison is that where the extended observational report (i.e. new facts are added) makes H false. We say that H is refuted by the experiment. This does *not* imply that H will be refuted again if the experiment is repeated. What is meant by a repetition of an experiment in the context of the mathematical game will be defined rigorously later.

Contrary to a refutation of H is a “realization” of H . What is meant by a realization of H ? We consider the following change of H for the extended observational report: we try to replace all the invented x_i of H by letters a_k of the extended observational report such that the new hypothesis is theoretically existent, i.e. is a theorem for MTA with A as the extended observational report. Then we say to have “realized” H . Not so rigorously we may say that H is realized if H becomes a part of the extended observational report.

Let us now ask: under what circumstances is it “possible” to realize H ?

Obviously the mathematical game itself cannot decide whether H can be realized or not. If PT is too weak, there can be many allowed hypotheses which cannot be realized. For instance, if we take thermodynamics without the second law, the mathematical game contains as allowed hypotheses so-called perpetuum-mobiles of the second kind. Thus we see again that the condition that a PT should show no contradictions between MT and the observational report, is too weak if we want to say whether a hypothesis can be realized. Without stronger requirements

on PT we cannot define what is meant by real and possible facts. We will not discuss these stronger requirements here (see LUDWIG 1985–1987: XIII, §4.6). We can avoid this discussion since we will only investigate the following special cases.

If H is theoretically existent no added laws can change the theoretical existence. H should be realizable since no new laws can forbid this realization. Therefore we say that H is “physically possible”.

But how is it with an allowed hypothesis H if there is another allowed H' which is not compatible with H ?

The observational report A can be taken as a hypothesis if we forget that it was read off from the facts. We write for this hypothesis H_A . A is a realization of H_A . Another realization of H_A is what we call a repetition of the experiment. Let H_0 be the hypothesis “ H_A and H ”. If this H_0 is theoretically existent there must be a realization of H_0 . But this does not imply that the given observational report A must have an extension which is a realization of H . On the contrary, it can be that such an extension gives a realization of H' which is not compatible with H . But we can repeat the experiment with another realization of H_A and perhaps a realization of H_0 , i.e. a realization of H .

In this situation we say that H is physically possible “before” the observational report A is extended. What do we mean by such a proposition? We just mean that H_0 is realizable, e.g. physically existent. To characterize this situation, we say: H is “conditionally physically possible”.

Let H be not only theoretically existent but also determined. The observational report can be extended in such a form that H can be realized by replacing the invented letters x_i of H by letters a_i of the extended observational report. Since H is determined, there cannot be any two different signs a_1 and a_2 by which one of the x_i can be replaced, i.e. for every x_i there can be only *one* fact in the fundamental domain corresponding to x_i . Since H is realizable there must exist one fact, even if we have not “reported” it, i.e. if the observational report does not contain the signs corresponding to this fact. (We have not reported it because it either lies in the future or we have not noted it.) Therefore we say that H is “physically real”.

To use only hypotheses of the first kind would make physics too clumsy. The fruitfulness of the physical language rests on the use of hypotheses of the first *and second* kind. One can transfer the interpretation language word by word to hypotheses of the second kind (see LUDWIG 1985–1987: XIII, §4.6). For instance, a theoretically existent

hypothesis is interpreted as physically possible and a theoretically existent and determined hypothesis as physically real. We say that by the observational report the new reality defined by the hypothesis is indirectly measured. The direct measurements are the observational reports.

The mathematical game with this physical interpretation of which we have given some examples becomes what we call the "physical game". We see that the interpretation language of "real" and "possible" in this physical game depends decisively on a classification of hypotheses, which is not a purely mathematical question in the scope of *MT*.

Nevertheless there are many structures in *MT* which are adjusted to this game and therefore interpreted by corresponding moves in the game. For instance an axiom of the form "there is an x for which . . . and so on . . ." is to be interpreted as: "it is possible to make an x for which . . . and so on . . .".

By the physical game we find new realities, for instance by the physical game of quantum mechanics the real atoms. We find also new possibilities as for instance atom bombs.

The interpretation language of the physical game can be systematically developed from the simple propositions of which we have given examples. One may introduce dialogue games with the intention to formulate a logic for this interpreting language. Decisions in this dialogue game are not only based on *MT* but also on the observational reports. It is not our intention to develop such a language and logic. Only one fundamental decision about this language and logic has already been made by our opinion of what physics is, shortly presented here: such a logic is a logic "*a posteriori*". It depends on the given structure of *MT*. The development of a *PT* needs only a primitive logic (characterized by words "and" and "not") for the formulation of the observational report and the mathematical logic of *MT*. There is no "new" logic "*a priori*" which determines the structure of *MT* and the formulation of the observational report. Some authors intend to develop a new logic "*a priori*" and to base quantum mechanics (i.e. some fundamental structures of *MT*) on this logic. From our point of view, this appears as if one were to construct bridles and by this construction try to prove the existence of horses for which the bridles are suitable.

The physical game demonstrates that there is no "pure" physics without technical applications. Such a pure physics would be a physics without the physical game, i.e. only *MT*, i.e. only pure mathematics.

Most of the observational reports describe realities produced by human actions, i.e. artifacts. The only "naturally" given and "very interesting"

observational report seems to come from astronomy; and even this is not given without indirect measurements by highly technical devices.

We as human beings are responsible how we play the physical game, e.g. what we realize or do not realize. Not all possibilities can be realized since there is not enough time and there are not enough human beings. We have to select what we want to realize.

No *MT* together with the correspondence rules can be evil. But moves in the physical game can indeed be evil, e.g. to make with human beings physical experiments which harm them. In many cases it is not simple to decide which move we ought to choose, since many circumstances must be taken into account. Thus different persons can reach different conclusions. It would be bad to suspect all who do not make the same decisions as we ourselves.

Reference

LUDWIG, G., 1985–1987, *An Axiomatic Basis for Quantum Mechanics* (2 vols.) (Springer, Berlin, Heidelberg, New York, Tokyo).

This Page Intentionally Left Blank

ON LEARNING FROM THE MISTAKES OF POSITIVISTS*

GRAHAM NERLICH

Department of Philosophy, University of Adelaide, Adelaide, South Australia 5001, Australia

1. Aims of the paper

Philosophers have much extended our understanding of the foundations of spacetime physics in this century. They go on doing so. Much of this is the work of positivists and conventionalists. FRIEDMAN (1983: Introduction) tells us in his excellent book that we must learn from their mistakes if we hope to go beyond them. I will look at two of their mistakes; one is about simultaneity and the other is about the relativity of motion.

In my view, the most interesting and suggestive arguments take the old-fashioned form that the world could not possibly be as our best scientific theories say it is. One argues that physics, as the scientists hand it down to us, is conceptually awry and must be rewritten, either by amputating parts of it or by interpreting these parts as conventions, not factual claims. They are metaphysical arguments. In this century, Mach, Einstein, Reichenbach and Grunbaum have urged them. Epistemology appears in them only by ruling out concepts from a fact stating role unless they meet some observational criterion or other.

It would be absurd to claim that this exhausts the normative role of epistemology in the philosophy of science. But I see little value in asking, about these dismembered forms of physical theory, whether they are better evidenced than the standard forms used by the practicing physicist. We philosophers are not likely to choose better than they do, even if we

*I am grateful to Graham Hall (Department of Mathematics, University of Aberdeen), Adrian Heathcote (University of Sydney) and Chris Mortensen (University of Adelaide) all of whom read the paper in some form and discussed it with me. They are not responsible for any mistakes in it.

enlarge their view of what they may choose among. Further, a main theme will be that a focus on the observability of ideas and on parsimony of theoretical structure misled us for decades as to what spacetime theories were about. Such preoccupations are still with us. I urge that we need to look, too, at which structures in a theory may help us grasp how it may evolve under the pressures of new evidence and speculation.

I make some large assumptions. No distinction between observation and theory statements will sustain the fashionable thesis that theory is underdetermined by all possible observation statements. I have argued elsewhere (NERLICH 1973) that Quine's pragmatist holism gives no workable account of the deep entrenchment of propositions in theory—the pragmatist surrogate of necessary truth—or of theory change. Also, that Quine confuses the range of observational vocabulary with the range of observable fact (NERLICH 1976a). I agree with others who have argued that no strong principle of charity is defensible. I believe that single sentences, as atoms of truth, have been neglected and ill understood by modern pragmatism. I have put these assumptions crudely and briefly, but the aim of my paper will be misunderstood without some brief reference to them.

2. Conventionalism and simultaneity

I turn first, and briefly, to the question of the allegedly conventional determination of simultaneity in SR.

BRIDGMAN's (1962) proposal ought to have closed the debate on the status of simultaneity. Use of slowly transported clocks properly defines a relation which is unique, relative and not arbitrary. This is clear in a spacetime picture; a glance shows that the definition can have no rival. Yet no consensus accepted it. Malament's analogous but more effective presentation of Robb's proof that we can define simultaneity in terms of causality has no advantage of greater transparency, nor of greater fidelity to real causal relations. It succeeded because it did make one, but only one, concession to conventionalism; it was explicitly causal. Thus the core of the objection to conventionalism was perfectly clear.

Let me illustrate how limpid Bridgman's definition is. Choose a reference frame and two rest clocks, *A* and *B*, at different places in it; represent them by two parallel timelike lines in spacetime (Fig. 1). Choose a point event, *e*, on one of them—*A*, for example. Imagine the

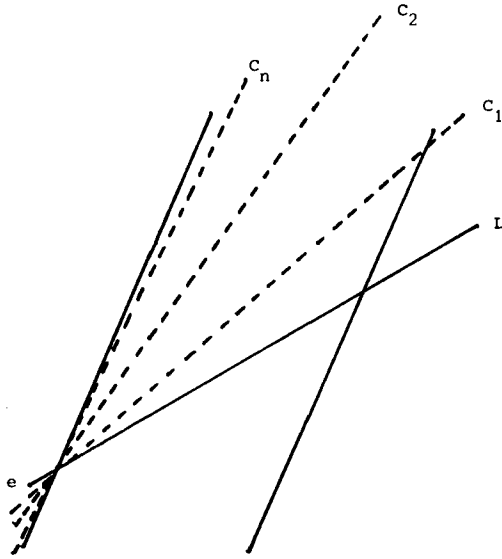


Fig. 1.

trajectories of uniformly moving clocks which might lie as close as you please to e and equally close to some point event on B . There are two limits to the trajectories of such clocks. First, there is the light ray which intersects e and some point event on B ; second, there are trajectories which approach arbitrarily close to parallelism with the worldlines of A and B while sharing a point with each. The first limit can offer no satisfactory definition of simultaneity, since different simultaneity relations emerge depending on whether we place e on A or on B (Fig. 2). We have no way of preferring A over B for the location of the reference event e . But if we choose parallelism with the frame clocks as our limit, then we get the same simultaneity relation whether we place e on A or on B .

Thus we have the ideal situation of a limiting trajectory which approaches as close as you please to A 's, as close as you please to B 's and which has a point in common with each. So long as we stick to the assumption that our clocks do, indeed, measure the length of their own worldlines, the definition is unique, non-arbitrary and relative, varying quite obviously with the direction of the worldlines (i.e. the frame of reference chosen). Once we accept the definition, then the trajectories in question are clearly those of clocks whose motion is arbitrarily slow relative to the chosen frame. Equally clearly, the resulting relation is

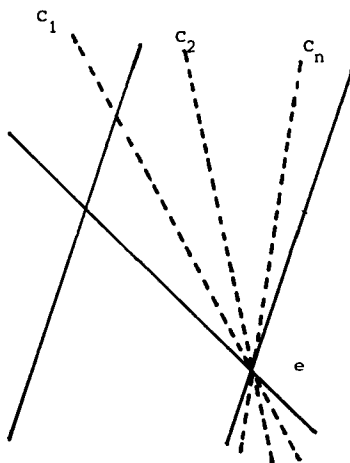


Fig. 2.

satisfactory in being neither absolute nor arbitrary. It is the same as Einstein's alleged convention.

If this is so clear, why was it not seen at once that Bridgman's suggestion settles the issue? Similarly, why was the work of Robb so long neglected? Partly, I suggest, because an argument about the issue was unlikely to be recognised unless it fell within the conventionalist-positivist problematic and style of presentation. Yet then, it was likely to be obscured by other conventionalist preoccupations. Bridgman, Ellis and Bowman were noticed because they argued inside that problematic. But it is foreign to the real thrust of the definition. A structural issue was obscured by too much attention to Ockhamist, reductionist ambitions. There are further reasons for seeing this episode as underlining that our thinking about SR was partly misled by Einstein and Reichenbach. And seriously misled.

A blindness to issues other than parsimony and observability obscured and confused the issue. It is not a simple matter of the scope of causal definition. Conventions for simultaneity can begin to make sense only in a particular—and imperfectly clear—conceptual setting: one where we speak seriously of the identification of places at different times and thus of the same time at different places. That is the language of frames of reference, not coordinate systems. Frames are distinct only if their rest points and their simultaneity classes differ, whereas coordinate systems differ more finely; for example, if we use polar rather than Cartesian coordinates for the same space of rest points. There is hardly

an issue of convention or of anything else about simultaneity unless we take frames as somehow significantly different from coordinate systems. Frames of reference give us physics understood not through spacetime, but through space and time. The issue gets to be of significance only if we give some weight to taking space and time by themselves. There are no relevant non-factual, merely conventional statements in a spacetime treatment. Yet any convention in the one case should appear in the other. Relative to a coordinate system, the question whether or not events have different *time* coordinates has just the same status as the question whether they have different *spatial* ones. Neither question is seriously about rest or simultaneity. The spacelike size of time coordinate slices of things is well defined in coordinate spacetime physics, whether or not their spatial coordinates are identical for different time coordinates. But that is not the same as the thing's spatial shape and size. None of these mismatches can licence a distinction between fact and convention as to simultaneity in a frame of reference of which there is no trace when we deal with coordinate systems. That is just one of several closely allied confusions.

Another is about the relation of frames of reference to coordinate systems for spacetime. That GR was written covariantly seemed to argue that it also lifts SRs restriction to the privileged set of Lorentz frames. But general covariance among coordinate systems is a much weaker condition than equality among frames of reference and complete symmetry for motion in all spacetimes — that depends on the structure of the spacetime itself. Clearly, inertial frames are preferred for flat spacetime. Someone might think that if skew coordinate systems are usable for a covariantly written theory, deviant relations of simultaneity in corresponding inertial frames must be permissible, too. But it is not so.

Second, the idea of a convention is the idea of a distinction between sentences which report facts and those which do not. It implies that real structure is less than we find in theoretical language. But the use of skew coordinates has nothing to do with that; it can only multiply the stock of things we relate factual descriptions to. In fact, use of skew reference frames gives an unacceptable space which is anisotropic in all sorts of ways. Even if it did not, it would not give us a *convention*. It is not clear what role conventions can play in simultaneity. Sometimes the literature reads as if it is telling us that moving bodies have no definite shape and size in the frame. It seems to say that, though rest and motion are matters of relative fact, simultaneity has no such factual status. It waits to be determined, even after all relative matters are fixed. That is confusion.

The source of this lies in Einstein's treatment of simultaneity, which *seems* to licence SR by purging some classical spatiotemporal structure. In fact, of course, it adds some. This could not have been understood before 1908 and the advent of spacetime. Certainly, SR removes the space stratification of classical spacetime, that is, its absolute partitioning into spaces at times. But SR replaces it with a metric of spacetime, whereas Newtonian spacetime, even if curved, has only affine structure. So SR has more structure, not less. Conventionalism really did obscure this and caused SR to be seriously misunderstood for quite a long time. There is another way to put this misunderstanding: it was not seen that the concept of reference frame, as we find it in classical physics, is inadequate for use in a spacetime with full metric structure. It needs revision. In classical physics, frame relative descriptions of physical events are completely fixed by a choice of the spacetime curves to represent the frame's rest points. But SR is a more structured theory than classical physics and choice of rest points does not give enough. Thus conventionalism told us precisely the opposite of the truth about the nature of this theory and its factual richness. It was false that a convention allowed use of familiar sentences, though a less structured world robbed them of factual import. Exactly the opposite is true; a factually richer world demanded a richer concept of reference frame to portray it fully. To describe all the richness of the world in frame relative facts¹ we need a reference frame which stated its simultaneity classes as well as its rest points!

3. The relativity of motion

The relativity of motion is still imperfectly understood, I think. People thought it ought to get into physics as some sort of axiom, being either analytic, *a priori* true or, at the very least, self-evidently desirable. It is easy to see why; if motion is change of place and place is a complex relation of distance and direction among bodies, then motion has to be a symmetrical relation among them. Other grounds were offered too, but epistemological demands on definitions for physical concepts have been among the more important. Many have felt that GR validates the relativity of motion in a way that meets the demands. So GR seems to

¹ The proviso about richness in frame relative facts is intended to cover, for example, the relative shape and size of moving things — their *spatial* characters rather than the *spacelike* dimensions of their coordinate cross sections. See also NERLICH (1982).

commend, to scientists and philosophers alike, a positivist approach to spacetime foundations.

The relativity of motion is a big step toward getting space out of our ontology. That looks desirable because space is obnoxious to ontology and epistemology alike. It does enough work in classical mechanics to be indispensable but not enough to be intelligible. It seems featureless and intangible to the point of vacuity. Leibniz's objection, that it could make no difference to have created the universe somewhere else in space with all spatial relations among bodies just the same, looked plainly unanswerable.

But this was delusive. There is nothing *a priori* about the relativity of motion. Only naivety about geometry makes it seem so. It ought never to have figured as an *a priori* requirement on physical theory.

Still, something is wrong in Newtonian physics, and classical physics generally. It has been much discussed. Newton makes an absolute distinction between bodies that accelerate and those that do not. It is a causal distinction within dynamics. The related distinction between rest and uniform motion has no dynamical role; but Newton still wished to draw it. Later classical physicists spoke of acceleration absolutely, but of rest and motion only relative to some privileged frames of reference. They allowed themselves to speak of *changing* velocity absolutely without making sense of *velocity* absolutely. This is not satisfactory.

To call it unsatisfactory is not to call it false. The world could have been as Newton described it. It could also have been as the post-Newtonians described it. Each is unsatisfactory in a different way, and neither way is epistemological. Each description leaves the world incoherent; it gives us distinctions in kinematics which link with no distinctions in dynamics. It does not matter that identity of place is unobservable, but it does matter that it has no causal role in physics, while being a physical distinction. What is physics about if not space, time and matter? To put the ugliness of classical physics another way, Galilean relativity gives us the sole classical example of change without cause; whether we see it as change of absolute place or change of distance and direction among bodies makes no difference. It is still an ugly affront to the classical presupposition that all changes are caused. The rest of classical science strongly vindicates this presupposition. The scandal of absolute space lies in that affront, not in its unobservability. But it gives a powerful motive for making the relativity of motion an *a priori* thesis.

However, rest and uniform motion are essentially observational concepts in a broadly classical setting; that is, one in which space and time

are considered as separate. (If you prefer, think of a spacetime stratified at each time into a unique S_3 space, for instance.) This could not have been understood in the 17th century for no one knew about non-Euclidean geometries then. Spaces of constant curvature, either positive or negative, offer free mobility, as Helmholtz called it. That is, an elastic solid which fills any region of the space in a tension-free state, can so occupy any other. But an object which is in a relaxed state while at rest will not be relaxed while it is moving, if it moves in a constantly curved space. It will move under tension and distort its rest shape as a function of its speed and its deformability under a given force. Uniform motion of elastic solids is measurable by strain gauges in non-Euclidean space.

This is easy to visualise in the case of the uniform motion of particles in a dust cloud through a space of constant positive curvature. (I am assuming, here, a classical relation between space and time, not some sort of curvature of a spacetime.) For there, all geodesics eventually intersect and then diverge. In Euclidean space a dust cloud may keep its shape and size just as well in motion as at rest, since the velocity vector of each particle may be the same, in magnitude and direction, as that of every other; so the particles can move along parallel trajectories at the same speed. But this possibility is unique to Euclidean geometry, among spaces of constant curvature, at least. In elliptic space, the geodesics which the bits of dust move along intersect. The result is just as observable if the curvature is negative. We can consistently suppose that Newton's laws still hold for particles in such a space. A cloud of free fall dust, in elliptic or hyperbolic space, can keep a volume constant in shape and size only if it stays at rest. Absolute rest is kinematically definable. Yet Newtonian laws apply to every point mass, so we here are envisaging worlds in which Galilean relativity still holds for them. The distinction has been made observable, yet still lacks a role in dynamics; *the breach of the causal principle is not filled*. This is still an incoherent world picture, ripe to create conceptual dissatisfaction.

Even in a world where we can never actually observe them, as in Newton's world, rest and uniform motion are in principle distinguishable by observation. The distinction passes rigid positivist and operationalist constraints. We can always raise the testable conjecture that space has some slight curvature to make practical the verifiability-in-principle of uniform motion. It was simply bad luck that the suggestion could not be made in the 17th century.

Variable space curvature yields stronger results. It cripples all those indiscernability arguments which Leibniz used so persuasively. For in

such spaces there is no isotropy and homogeneity from one point to another; spatial points differ in the structures that surround them, in purely spatial ways. So it really might make a difference if all the matter in the universe was one metre to the left of where it actually is. It could be that only that allows all matter to fit into volumes whose geometry lets them be relaxed there. There is no free mobility, even in Helmholtz's sense, if the curvature of space may vary.²

4. Non-decisive *a priori* criteria

Let us consider a Newtonian stratified spacetime, in which we make no appeal to curvature to account for gravitation. It offers a new perspective on the incoherence. This flat Newtonian spacetime is like a pile of hyperspatial sheets, pierced, as if with spokes, by time extended particles. The sheets may slide across each other, just as a stack of paper sheets would do, taking the spokes with them and retaining the places where they pierce the individual sheets. Each paper sheet has its own Euclidean spatial metric; the thickness of the sheets gives a time metric. The pile fills a Euclidean 4-space in which all force free trajectories are straight. Nothing in the model corresponds to the length of a spoke (i.e. of a particle worldline). Nor is there a metric along the curved spokes of accelerating particles, of course. So nothing corresponds, either, to the orthogonality of any straight trajectory to the spacetime sheets they pierce. If we could speak of a spoke as piercing some n seconds of stacked sheets with a minimum spacetime length, then the spacetime would define orthogonal piercing and would define points of absolute rest and thus of absolute uniform motion. These concepts would be definite whether or not our perceptual powers equipped us to detect them. But the affine structure of Newtonian spacetime makes no such distinction. Nor does it help to embed the pile of sheets in a curved spacetime where all free fall particles are geodesics and gravity is caught up in geometry. So, again, it is incoherent, conceptually: the product structure of the

² Note that I am not talking about GR here, where the indiscernability arguments may be rescued. I mean a classical stratified structure, e.g. a constant S_3 at each time. In GR, the curvature of space accompanies the distribution of matter, so there is always another possible world like the one in question, but with all the matter and all the curvature displaced together in new regions. But if we imagine a possible world where curvature is not tied to matter distribution, this salvation fails (SKLAR 1985: p. 14.).

spacetime yields both metric space and metric time, without any metric in the spacetime itself.

It is not inconceivable that the world should be this way, of course. But, I claim, this is *non-decisively* objectionable on *a priori*, conceptual grounds; that is, in such a world, scientists and philosophers would have legitimate grounds for seeking a more coherent conceptual structure, even if the world were just as this theory describes it. A true account of the world might leave us dissatisfied, with legitimate, specifiable, though not necessitating, grounds for trying to improve the account. In such a world there is a proper case for trying to reconceive it more coherently. We might try and fail, but not because we were obtuse. I am proposing *a priori* reasons which do not necessitate, but which appeal to criteria of coherence. This is something like the converse of Kripke's suggestion that there may be necessary propositions of identity which are not knowable *a priori*. Had Newton been right, I suspect that debates about space would have persisted unresolved in epistemology, physics and metaphysics. For, I think, space is not reducible to spatial relations among objects, yet its role in Newton's physics is obscure and unsatisfactory. That is, Newton described a quite possible, but imperfectly coherent, world.

I call these principles *a priori* because they *lead* theory choice; they are *reasons* for a choice. Principles of teleology, determinism, continuity and the principle that physics can be truly expressed in laws, are among them. They are not necessary, since teleology, determinism and continuity are either false or presently improbable. The principle that there are true laws is deeply entrenched, despite our belief that we do not yet know any law which is precisely correct. But it is not necessary. That is one way in which these principles are not decisive reasons for choosing theories.

I have tried to illustrate another. Whether such a principle is true or false may depend on the success or failure of some philosophical programme of reduction. I assume that either Newton's version or a post-Newtonian version of classical physics could be true, but not both. I also assume that not both a relationist and a substantialist version of these theories can be true. I do not think that a relationist reduction is possible, but a classical world would always present us with a good motive for trying to make it work, since it is an incoherent world: it disjoins uniform motion and mechanics. We still do not quite know how to settle these factual philosophical debates. Nor are we likely to find out until we learn more about how to identify and describe what I am here calling non-decisive *a priori* reasons round which the debate centres.

Principles of observation and coherence overlap. Any fundamental

physical theory applies to every physical entity; thus every basic theory has the same domain as every other, though not the same ideology, of course. It follows that a theory will apply to measuring instruments and to observers and to processes within them involving the properties which the theory is about. The relation between the vocabulary of the theory and that of flashes, bangs, smells and tastes can only be as vague and unstructured as the latter vocabulary itself. This point hardly makes sense save within a realist context in which every subtheory in physics applies to a single domain of theoretical entities.

It may be, of course, that principles about the relation of scientific concepts to observational ones are among these *a priori* reasons which do not necessitate.

For these reasons, the supposed advantages of an unrestricted relativity of motion were illusory. On the one hand, rest and motion are, in principle, quite as observable and absolute as acceleration is. On the other, SR, in providing a metric for spacetime, provided the first metaphysically coherent and intelligible arena for physical events, in which, although absolute rest was not defined, its metaphysics gave no presumption that it should be well defined. It commits one, not to space and places, but to spacetime. Change of place is not a basic idea of the theory because space itself is not. The crucial distinction lies between linear and curved trajectories in spacetime. No way to identify places across time is needed since space and time, by themselves, have become mere shadows.

5. The geometry of force

Another thread woven into the fabric of the relativity of motion is an idea about force; that the spatial relation between the force's source and a test body which it acts on should be rather simple. Classically, forces meet something like the following criteria (NERLICH 1976b: p. 218).

- (i) Any force has an identifiable body as its source to which its target is spatially related in a definite way. Sources are force centres.
- (ii) The conditions under which the source body acts are specifiable independently of any description of its effect (e.g. a glass rod is electrically charged when rubbed with a silk cloth).
- (iii) Each centre of force acts on its target in some definite law governed way (by contact, inversely proportional to the square of the distance).

- (iv) Generalisations describing the action of forces (e.g. all like charged bodies repel) are defeasible (e.g. unless there is an insulating wall between them) but the defeating conditions must be causal.

Criterion (iii) works in classical mechanics like this. In an inertial frame, any test body whose spatial path is not a straight line can be linked to a source object, so that any force vectors which curve the body's path, are related rather simply to the position vectors which point at the source. Even when the vectors of magnetic force round a moving electron are found to be orthogonal to the position vector linking a test body to the source, and in a plane orthogonal to the motion of the source, the relation is simple enough to let us see the field as an emanation of the source. For non-inertial frames, the force vectors for centrifugal and inertial forces have no such simple relation to sources. Indeed, that is plain from the stronger fact that the centrifugal and inertial force vectors will be the same whatever other objects there may or may not be elsewhere. But, I submit, it is only if the underlying geometry is Euclidean that the more general criterion linking force vectors geometrically to source position vectors is swallowed by the stronger criterion that real force vectors must depend *somehow* on the presence of sources. Only Euclidean geometry can realise the stronger, simpler criterion.

This use of criterion (iii) is important for relationism and spatial reduction. That is because it gives a prominent, *physical* role to spatial relations among objects. The position vectors are spatial relations based firmly in dynamics; that plausibly contrasts with the wider embedding space which appears as a cloud of merely possible force and position vector links. I do not think this attraction overrides other difficulties which have been found in relationism, but I conjecture that the way it stresses spatial relations plays some role in reductive thinking.

There is some evidence of this in Mach's critique of Newton's rotating bucket experiment. Though Mach complained that Newton's theory conjures with unobservables, the real physics of his proposed answer to Newton's challenge makes the fixed stars sources of the inertial field. The field forces were envisaged as meeting the criteria just set out. Mach's proposal links centrifugal force vectors to position vectors for the stars in a way that was sufficiently simple, given the symmetry of the star shell. The relative motion of the symmetric shell through the non-inertial frame produces forces on the objects at rest in the frame. These force vectors at least *begin* to relate, feasibly, to symmetric position vectors and the

direction of their change. The *a priori* complaint about unobservable space cannot be central, because the revision proposed results in more than conceptual changes. I am suggesting that Mach's real wish was to avoid the theoretical incoherencies that I am trying to illustrate and identify.

6. The relativity of motion in general relativity

Whether the unrestricted relativity of motion is a thesis of GR in any philosophically interesting sense depends, as before, on our willingness to take rather seriously a distinction between frames of reference and coordinate systems. The distinction is, arguably, not clear enough to allow us to do so. If we do the physics of SR through reference frames, then "accelerating" objects can no more function as frames of reference than they can in classical physics. But this says little about our willingness to work within the corresponding spacetime coordinate systems. So it is not perfectly clear that the sort of question that occupied the positivists about the relativity of motion is not simply left behind in GR. But I hope some light may still be shed on philosophical issues by treating frames in a serious, quasi-classical way.

Why does the unrestricted relativity of motion fail for flat spacetimes? Firstly, because the very strong symmetries of flat spacetimes allow us to pick an infinitely large set of linear time-orthogonal coordinate systems, rather simply related to each other by the Lorentz transformation. To these there correspond rigid and global frames of reference, simply transformable into each other through relations which have the dimensions of a velocity. They are privileged frames in flat spacetime, because any other consistent frames of reference are either merely local, have anisotropic spaces, or have spatial and temporal metrics which change from place to place and time to time. So only a restricted relativity of motion holds here.

Nevertheless, though flatness reveals its strong flavour in this way, it produces Euclidean spaces (or everywhere locally Euclidean ones, e.g. in hypertoroid spaces) which, as I already mentioned, strike us as bland to the point of vacuity. It allows free mobility in Helmholtz's sense and it fails to distinguish kinematically between rest and uniform motion. So the very features which make flat spacetime yield a space which looks like a nothing, and thus ripe for reduction, also deliver an unequivocal preference for highly symmetrical and global reference frames. The preference

restricts the relativity of motion, thus imperilling the reductionist programme.

We get the opposite state of affairs, in general, when we move to the more complex, variably curved spacetimes of GR. We may lose any or all of the criteria which make Lorentz frames desirable in flat spacetimes. To start with, there may be no possibility of global reference frames. But even where we can use them, we will find, in general, that they are not rigid; clocks will run at different rates in different places and space expand or contract as time goes on. That is to say, we will be obliged to treat the *coordinate* differences in time and distance seriously if we are taking the idea of frame relative motion seriously. Of course, arbitrary choice of frame of reference may exaggerate this lack of rigidity wildly. But, in variably curved spacetimes, we lose a general contrast between rigid and non-rigid frames, and between frames that do and those that do not give us isotropic spaces.

Nevertheless, the same variable spacetime geometry, which swallows up preferred frames in its asymmetries, also, in general, gives the spaces of these arbitrary frames an obtrusive and changing geometry which has a very distinctive and prominent kinematic and dynamic character. It is impossible to think of a space as a featureless nothing, if voluminous solids are stressed, or even shattered, simply by moving inertially into regions where the geometry leaves no room for them to exist undistorted.

Consider the problem of an arbitrarily selected local frame in any spacetime, where we treat the space and time of the frame as given by the set of t -constant spacelike hypersurfaces and x_1 - x_3 -constant timelike lines that make up the coordinate lines and planes of some arbitrary coordinate system. The only requirement on the various, arbitrarily curved, spacelike hyperplanes is that they must not intersect; similarly for the arbitrary timelike curves. When we project down from this arbitrary coordinate system into the space and time of the frame, the resulting arbitrary spatial geometry will have geodesics of its own which bear no simple relation to the geodesics of the spacetime (as it is easy to see from the special case of flat spacetimes). The motion of any free fall particle whose spatial path in the frame is not geodesical, will be guided on its curve by vectors of the gravitational field, as determined by projection from the spacetime curvature of the region. These vectors can certainly not be expected to link, in any simple geometrical way, with position vectors, even of nearby sources of the matter tensor and the curvature guidance field, let alone of the array of sources as a whole. This will be especially true, of course, when the geometry of the frame's space is itself

complex and changing. But even if it is not, we lose any simple link between force and position vectors. I have suggested that this gives most of its point to the focus on spatial relations, so prominent in reductionist literature.

Finally, GR is rather rich in spacetimes which yield preferred reference frames (or classes of frames), both globally and locally. Nor are such examples among the rarer and more exotic models. Flat spacetimes are examples, as is the spacetime model of our universe preferred in modern cosmology. In fact Rosen has suggested, in several papers (e.g. ROSEN 1980), that GR might be rewritten in such a way as to make this fact formally more prominent. This has not been widely accepted, but it serves to emphasise the unintended emergence of preferred frames.

It remains true that the basic concepts of GR are relativistically forged in that the theory is written covariantly in a non-trivial way. But that does not mean that the unrestricted relativity of motion is a key concept of the theory, nor that the best way for us to learn about the metaphysics of spacetime structure is to understand the mistakes of the positivists nor the philosophical relativists more widely. Their ideas were often confused and baseless.

7. Other ways of understanding spacetime

Some think that only conventionalism and positivism offer any systematic picture of the foundations of spacetime theories (SKLAR 1985: p. 303). But this is not so. Conventionalism avowedly fails to find any foundation at all for large parts of the theories, writing them off as mere non-factual conventions. Where it does offer foundations they are narrow, opaque or dubious, as in the case of causality, or confused as in the case of simultaneity. A more pluralistic view of the search for foundations is needed. We need a metaphor of foundations without *foundationalism* — at least, we do if we are to take seriously the title of this section of the Conference.

Let me turn now to describe three ways, quite different from each other, in which we might properly regard a philosophical study (or a quasi-philosophical one) as tracing the foundations of spacetime theory. These by no means exhaust the range of options and are not even intended to represent the traditional main stream of what might be called *foundational studies*.

Positivist and conventionalist attempts to find the foundations of

physical theories are motivated, clearly enough, by the laudably modest aim of saying no more than one must. Yet they also have an immodest tendency to rob theories of legitimate assertive power.

A taste for modesty may lead us to substitute a particular, favoured expression for other extensionally equivalent but intensionally different expressions. Relativity theories provide several striking examples of this; most notably, the very common reference to the null cone as the light cone. The word “causal” is very widely used for the relation, whatever it is, that divides the surface and interior of the cone from what lies beyond it. But, where positivism hardens its heart against intensional distinctions among coextensive expressions, its drive for economy is no longer simply modest. This is clear when one contrasts it with a modesty in realist attempts to arrive at a foundation for theory. I turn to that in a moment.

In general, positivist investigations search for what I shall call the *restrictive* foundations of theories. That consists of a minimal ideology — *foundationalism* in short — and a minimal set of axioms, which generate a body of theorems previously judged as indispensable. To call this the restrictive foundation stresses its tendency to reduce content and prune ideology. This is a bald account of positivist aims, to be sure, but I hope their familiarity will allow me to be brief and turn to something less familiar.

One motive behind a realist examination of a theory, is to look for what I shall call its *permissive* foundations. Whereas the restrictive foundation presents, in few axioms and a lean stock of predicates, all the theorems we simply must have, the permissive foundation presents axioms and predicates within which we can speculate most radically on how that same theory may develop. We can reflect on the results of dropping quite deeply entrenched propositions from a theory. Radical speculation can legitimately rest on the theory’s broader base for its development. The permissive foundations tell us what this base is, and which theorems might yield to correction without the theory’s being abandoned. It is not clear that these restrictive and permissive motives for finding foundations must lead in different directions, but in spacetime theory they certainly do.

It is very widely believed that a single principle about causality lies at the foundations of the relativity theories. It is the Limit Principle, that nothing outstrips light. The nice things about this Principle are that it is, very likely, true, that it is qualitative, and expressed in lean and intuitive ideology — simply “particle” and “outstrips”. If we take this as giving the core of the theory, we are likely to include it among the axioms, and to

allow ourselves to speak of the null cone as the light cone, since photons will lie in the conical surface. That is its restrictive foundation.

But there is a price for this. It may cripple speculation within the theory. Questions which might be fruitful cannot be pursued as developments of *that* theory, nor call on its resources, to make speculation clear and definite. The hypotheses that there are tachyons, and that the photon has some finite mass, cannot be pursued within SR if we formalise it according to its restrictive foundation, yet there is a clear sense in which both these suggestions have been investigated consistently within SR.³ Unless each was compatible with the foundations of SR, in some clear sense, neither could have been considered. The idea of permissive foundations for a theory may give to realism a theoretical modesty of its own; for it is clear that there may be a proper diffidence about denying structure just as there is about asserting it. In the present case, suppose that we admit null cones, which are observationally remote structures in an equally remote object, spacetime; then a realistic attitude towards them lets us speculate about epistemically more proximate objects, tachyons and photons, in a tangible and articulate form given by the admission.

When we ask for a permissive foundation, the question is quite differently motivated, and our criteria for a good answer are not at all the same, as for a restrictive foundation. Though the two hypotheses just mentioned are improbable, we can usefully ask how SR would survive their truth. The answer is quite obvious from a glance at tachyon theory, for all of its equations are Lorentz invariant. So the permissive foundation of SR is the Invariance Principle, not the Limit Principle. The Invariance Principle simple refers to the conical structure within spacetime and to the very powerful symmetries of Minkowski spacetime. The hypothesis of the massive photon may call for some changes in the laws of electromagnetism and optics, but whatever laws are suggested are still required to be Lorentz invariant. So that Principle is the permissive core of SR.⁴

Third and lastly, I want to discuss Ehlers, Pirani and Schild's well-known paper on the foundations of GR. In this, the authors link various geometrical structures—projective, conformal, affine and so on—to material structures such as photon and particle free fall trajectories. In what way is this paper foundational?

³ See FEINBERG (1967) and GOLDHABER and NIETO (1971).

⁴ See also NERLICH and WESTWELL-ROPER (1985).

Two physical conditions are needed for it to work at all; particles in free fall and null cone surfaces filled by light. However, GR is not a theory about the constitution of matter and has no ontic commitment to matter in particulate form. It can admit tachyons, and speculation about massive photons. For these reasons, we cannot see E, P & S as presenting an ontology, nor even a likely epistemology, in any standard sense, since neither particles nor the geometry of light propagation is observationally proximate. Yet something like each of these studies is at stake in the paper.

Free fall *particles* are needed because only point masses can be relied on to trace out the geodesics of spacetime which fall inside the null cone. These trajectories constitute the projective structure of spacetime. The centres of gravity of voluminous, elastic, massive bodies will not do, since the geodesics through various points in the solid will usually not be parallel. The bodies are gravitational multipoles. Internal stresses will tend to force the falling body off geodesical trajectories, even in the absence of external forces. The worldline of neither its geometric centre, nor its centre of gravity will be a geodesic. So, in general, it is only freely falling *particles* which inscribe geodesics. Similar reflections apply to the filling of the null cone. Of course the surface of the null cone will still be the boundary between spacelike and timelike curves; it will be the overlap of points elsewhere and elsewhen from the apex of the cone whether or not light or matter fills it. So what E, P & S offer us is an elegant and familiar way in which the geometric structure of GR might be *inscribed*. They single out a subset of worlds, which the theory makes possible, in which the geometric structures at the core of the theory and matter structures, familiar from a range of other theories, come together so that the latter trace out the former. This might be called an investigation of the inscriptional foundations of a theory, presenting one possibility within it.

I close with the suggestion that the philosophical investigation of physical theories may take several useful forms, that the form of a theory that philosophers find in scientific use is likely to be irreducible, and that time spent in attempting to reduce it to something else may illuminate why the theory is composed as it is, but will seldom result in a justifiable revision of it. Of course investigations into the epistemology of a theory shed light on what goes on in it, but epistemic theories of concept formation have seldom proved constructive or even insightful. One can usefully recognise *a priori* elements in a theory without relegating them to mere convention, and hope, eventually, to see how they are corrigible by

observation. In short, there is a variety of ways in which we might look for foundations in physical theory, for many of which it is simply unhelpful to complain that they raise epistemological problems which they make no offer to solve.

References

- BRIDGMAN, P.W., 1962, *A Sophisticate's Primer of Relativity* (Middletown, Conn.).
- EHLERS, J., PIRANI, F.A.E. and SCHILD, A., *The geometry of free fall and light propagation*: in: L. O'Raiheartaigh, ed., *General Relativity* (Oxford), pp. 63–84.
- ELLIS, B.D. and BOWMAN, P., 1967, *Conventionality in distant simultaneity*. *Philosophy of Science* 34, pp. 116–136.
- FEINBERG, G., 1967, *Possibility of faster-than-light particles*. *Physical Review* 159, pp. 1089–1105.
- FRIEDMAN, M., 1983, *Foundations of Space-Time Theories: Relativistic Physics and Philosophy of Science* (Princeton University Press).
- GOLDHABER, A. and NIETO, M., 1971, *Terrestrial and extraterrestrial limits on the photon mass*. *Reviews of Modern Physics* 43, pp. 277–295.
- MALAMENT, D., 1977, *Causal theories of time and the conventionality of simultaneity*. *Nous* 11, pp. 293–300.
- NERLICH, G., 1973, *Pragmatically necessary statements*. *Nous* 7, pp. 247–268.
- NERLICH, G., 1976a, *Quine's 'real ground'*. *Analysis* 37, pp. 15–19.
- NERLICH, G., 1976b, *The Shape of Space* (Cambridge).
- NERLICH, G., 1982, *Simultaneity and convention in special relativity*, in: R. McLaughlin, ed., *What? Where? When? Why?* (Reidel).
- NERLICH, G. and WESTWELL-ROPER, A., 1985, *What ontology can be about*. *Australasian Journal of Philosophy* 63, pp. 127–142.
- ROSEN, N., 1980, *Bimetric General Relativity and Cosmology*. *General Relativity and Gravitation*, 12, pp. 493–510.
- SKLAR, L., 1985, *Philosophy and Spacetime Physics* (California University Press).

This Page Intentionally Left Blank

9
Foundations of
Biological Sciences

This Page Intentionally Left Blank

EVOLUTION — MATTER OF FACT OR METAPHYSICAL IDEA?

ROLF LÖTHER

Akademie der Wissenschaften der DDR, Zentralinstitut für Philosophie, 1086 Berlin, GDR

How do we know that organismic evolution indeed has occurred in the past and goes on? We cannot observe evolution immediately at the present time. Although evolution is taking place today, it is much too slow to allow us to recognize evolutionary processes as such if we do not know by other means that they are evolutionary processes. The occurrence of evolution is demonstrated not by direct observation but in other ways. How?

In many textbooks of evolutionary biology we find chapters containing so-called evidence for evolution from different parts of biology: from systematics, comparative anatomy, comparative embryology, serology, comparative biochemistry, biogeography, paleontology, and so on. Essential to this kind of demonstration are the homologous similarities of organismic structures. The anti-evolutionistic objection that evolutionary biology is founded on a *circulus vitiosus* is related to this kind of demonstration: on the one hand there is the conclusion from the patterns of homologous similarity to evolution, on the other hand there is the explanation of the origin of the same patterns by evolution (cf. KUHN 1947, RIEPPEL 1983). That, indeed, is a *circulus vitiosus*. But this objection is not evidence against evolutionary biology. It is only an objection against an incorrect argumentation.

That evolution has taken place and takes place is demonstrated by the theory of descent which is fundamental to evolutionary biology (cp. LÖTHER 1972, 1983, TSCHULOK 1922). The theory of descent presupposes the conceptual reflection of the gradual diversity of organisms and is the explanation of the gradual organismic diversity. The patterns of homologous similarities are aspects of the organismic diversity which is to be explained. The theory of descent explains organismic diversity by the gradual descent of recent organisms (species) from common ancestors

(ancestral species) which lived before them, and by their evolution along divergent paths.

This is an inescapable consequence of the combination of the natural classification of organisms with certain general statements of biology. This combination leads to the conclusion that the gradual diversity of organisms is the result of descent and evolution along divergent paths. The truth of this explanation depends on the truth of the general statements. The general statements in question are

- the continuity of life on earth by means of reproduction since its abiogenetic origin from non-living matter (the principle of Redi “*omne vivum e vivo*”);
- the continuity of the specific organization of the living beings that is founded on the transmission of the genetic information in the process of reproduction;
- the genetic variability of the organisms which results ultimately from mutations;
- a process by means of which the genetic differences of the organismic individuals become selected, that is, eliminated or stored up, accumulated, and combined in a manner that may — dependent on time in the series of generations — result in great differences in the organization of living beings.

These four statements are proved by the development of biology. They are proved by the refutation of the conceptions of recent spontaneous generation from F. Redi to L. Pasteur, by the cognition of the reproduction of life since the statement “*omnis cellula e cellula*” (F.V. Raspail, R. Virchow), by the development of genetics from Mendel to the cognition of the regularities of the replication of DNA, of genetic transcription and translation and of mutability, and last but not least by Darwin’s discovery of natural selection in the struggle for life.

The relation of natural classification system of the organisms (species) and the theory of descent in biology has some parallels in the relations of other natural classification systems and theories or hypotheses which explain the phenomena shown by natural classification. There are, for instance, the relations of the classification of crystals according to their planes of symmetry, the periodic system of chemical elements, the classification of stars according to their place in the Hertzsprung-Russell-diagram, and the corresponding explanatory theories or hypotheses. In the case of the natural system of organisms the explanation is the theory

of descent, in the case of the periodic system of chemical elements it is quantum mechanics. The general pattern is

$$\frac{\text{classificandum}}{\text{classificans}} \rightarrow \text{classificatum} = \frac{\text{explanandum}}{\text{explanans}} \rightarrow \text{explanatio} .$$

As the theory of descent explains the recent diversity of organisms by descent and evolution in the past, it opens the approach of evolutionary biology to evolution as its field of research. Organismic evolution is an objective reality that is demonstrated in a theoretical way by logical conclusions from empirical statements of facts. The approach to evolution is mediated by the theory of descent.

The fossils are not part of the demonstration of the theory of descent. The theory of descent is independent of the proofs of the history of life in the past. In connection with the ideal reconstruction of the history of the earth by the geological sciences, the fossils provide additional confirmation for the theory of descent. The existence of fossils in the strata of the supercrust of the earth leads after all to the same question as the recent gradual diversity of the organisms, and with that to the theory of descent. This presupposes the identification of the fossils as traces of former life and their comparison with recent life.

Independently from evolutionary biology and in a logically similar way, historical geology opens its way to the historicity of the earth and by that its subject also to the study of organismic evolution in space and time. Palaeontology has its place between historical geology and evolutionary biology, giving to historical geology indicators of the relative chronology of the history of the earth, and at the same time material to evolutionary biology for the ideal reconstruction of the history of life.

Two questions follow from the theory of descent: (1) What are the concrete paths and modes of evolution in the past? (2) What are the factors, moving forces, and regularities of evolution? The first question gets its answer by historical phylogenetics. It leads to the ideal reconstruction of evolution in the past. The concentrated results in this area are graphically expressed in phylogenetic trees (dendrograms). In connection with that, natural systematics is developed into phylogenetic systematics (W. Hennig). The answer to the second question is the theory of evolution. The theories of descent and of evolution were founded by Darwin, historical phylogenetics by E. Haeckel. Historical phylogenetics and the theory of evolution both presuppose the theory of descent, and illuminate one another. Through the coherence of the theory of descent,

historical phylogenetics, and the theory of evolution the questions of the fact, the course, and the factors, mechanisms, and moving forces of evolution become answered by means of the historical method, which was introduced into biology by Darwin.

The historical method is the general scientific method by which it is possible to recognize developmental processes insofar as they belong to the past. By the application of this method, the evolution of the celestial bodies becomes evident in astronomy, the evolution of the earth in geology, the evolution of life in biology, and the history of mankind in the social sciences.

Historical research in natural sciences as well as in social sciences starts from the recognition of the present reality. According to the methodological principle of historicism, present reality is explained as a result of processes which happened in the past. According to the methodological principle of actualism, these processes are explained by causes which can be stated in the present time. The developmental theories in science originate in this way.

The historical recognition of evolution presupposes objectively that the present reality contains the past: not as a temporal succession of the past things, phenomena and processes, of their rise and decline, but in the co-existence and relations within the structures of the observable present reality. The material world is full of traces of its past — one must only be able to read them — just as it is full of the germs of newly emerging processes, of developmental processes coming into being. The present and future development of things and phenomena depends on the past inherent in them.

The way from the present co-existence to the succession in the past goes beyond the comparison of the things and phenomena and their classification according to their natural order. Thus the spectral classes and the luminosity of the stars lead to conceptions of their evolution, and information about the succession of the strata in the supercrust of the earth and the homologous structures of fossils and recent living beings are used to reconstruct their phylogeny. Karl Marx proceeded in the same way in the social sciences. He discovered the socio-economic formation as the unit of classification of the various human societies and of their place in the development of mankind. The term “formation” he obviously took from geology.

In comparison with the methods of empirical research (for instance observation and experiment), the historical method is composed of a complicated system of methods. It includes methods such as observation,

experiment, model-building, analysis and synthesis and so on. The structure of this system is determined by the two methodological principles of historicism and actualism. They are interconnected and they complement one another. In the process of the cognition of historical development they lead to retrodiction, the historical counterpart of the prediction of future processes. Historicism and actualism together enable the ideal reconstruction of the past on the basis of the present reality. While organismic evolution is stated as a matter of fact by the theory of descent, the cognition of the course, the ways and modes of evolution as well as of its factors, mechanisms, and moving forces proceeds through the interconnections of objective, absolute and relative truth.

The answer to the question how evolution has occurred, begins with the natural classification system of organisms. The explanation of the gradual diversity of living beings by the theory of descent leads to the following conclusion: Systematic relationships among the taxa by way of their places in the natural classification system essentially express real phylogenetic, genealogical relationships. They are reflected at each particular stage of the development of systematics more or less adequately. As a methodological consequence of this result the natural classification system used in systematics should be developed further into the phylogenetic system which represents a higher stage of its development, and — beginning with the analysis of the phylogenetic relationships of the recent species — reflects the phylogenetic relationships of the taxa in the classification system.

But as you know, there are some divergent schools in systematics, and vehement debates between them. The main schools are the traditional (classical, evolutionary) school, the phylogenetic (cladistic) school, and the phenetic school. Their differences are related to phenetic overall similarity and phylogenetic branching. There is also “transformed cladistics”, in my opinion a perversion of phylogenetic systematics; phylogenetic systematics without phylogenetics, which can be neglected in this context. With regard to the traditional, phylogenetic, and the phenetic school and the debates between them, the problem, as stated by GOULD (1984: p. 263), “arises from the complexity of the world, not from the fuzziness of human thought (although woolliness has made its usual contribution as well)”.

GOULD (1984: p. 365) reminds us of the source of the debate — “a rather simple point that somehow got lost in the heat. In an ideal world, there would be no conflict among the three schools — cladistics, phenetics, and the traditional school — and all would produce the same

classification for a given set of organisms. In this pipe-dream world, we would find a perfect correlation between phenetic similarity and recency of common ancestry (branching order); that is, the longer ago two groups of organisms are separated from a common ancestor, the more unlike they would now be in appearance and biological role. Cladists would establish an order of branching in time by shared derived characters. Pheneticists would crunch their numerous measures of similarity in their favorite computers and find the same order because the most dissimilar creatures would have the most ancient common ancestors. Traditionalists, finding complete congruence between their two sources of information, would join the chorused harmony of agreement."

Of course, we do not live in such an ideal world. There is no unambiguous correlation between cladogenesis and phenetic evolution. But is there nevertheless a possibility to overcome the conflict between mutually exclusive schools in systematics by a unified theory and methodology of systematics? As I see it, the approaches of phenetic, traditional, and cladistic systematics can be understood as stages of an advancing process of cognition. This becomes evident in the change of meaning of the term "systematical relationship" of the taxa among the three schools of systematics. SOKAL and SNEATH (1963: pp. 3/4) remarked: "There may be confusion over the term 'relationship'. This may imply relationship by ancestry . . . , or it may simply indicate the overall similarity as judged by the characters of the organisms without any implication as to their relationship by ancestry. For this meaning of overall similarity we have used the term 'affinity', which was in common use in pre-Darwinian times. We may also distinguish this sort of relationship from relationship by ancestry by calling it *phenetic relationship*, . . . to indicate that it is judged from the phenotype of the organism and not from its phylogeny."

Further phyletic relationships may be divided into two kinds. Two forms may be said to be closely related phyletically because they possess many characters which are derived from a common ancestor. The component of phyletic affinity which is due to such common ancestry (and not to convergence) is called *patristic* affinity. Second, the forms may be related closely through the recency of common ancestry, without taking account of the number of characters derived from a common ancestor. This relationship in terms of phyletic lines is called *cladistic* affinity. A cladistic relationship refers to the paths of the ancestral lineages and therefore describes the sequence of the branching of ancestral lines; it ignores evolutionary rates and is therefore not related to phenetic similarity. The

two aspects of a phyletic relationship cannot be considered to be additive. SOKAL and SNEATH (1963: p. 222) summarize:

“Phenetic Relationship = Homologous (patristic) + Homoplastic (convergent) Relationships

Phyletic Relationship presents two aspects:

- (1) Patristic Relationship
- (2) Cladistic Relationship.”

Phenetic, patristic, and cladistic relationships are concretisations of systematic relationships in the case of phenetic, traditional, and cladistic systematics. The transition from a phenetic relationship to a patristic relationship takes place by the recognition of the homologies, the transition from a patristic relationship to a cladistic relationship by the differentiation among the homologies between plesiomorph and apomorph characters. At both stages the monophyletic origin of the taxa is required. The transition to the cladistic stage involves a new, more precise concept of monophyletic descent. In traditional systematics, monophyletic descent only implies the derivation of a taxon over one or several consecutive ancestral forms from a direct ancestral form of the same or a lower systematic position. Compared to it, cladistics demands that all species of each higher taxon are derivable from a common ancestral species, and no species which descended from that ancestral species is allowed to be outside of this taxon.

In the transition from phenetic to patristic to cladistic relationships the strata of phenomena are pulled down to the level of the realization of a fundamental law of living nature — Redi’s principle “*omne vivum e vivo*”, which has its place in the explanation of the gradual diversity of organisms by the theory of descent. The relation of this principle to the phylogenetic (cladistic) system is that of a law of nature to the idealized description of a natural process which is a realization of the law in question. The cladistic system is an idealized representation of the natural process of evolution with regard to the principle of Redi and seen from the temporal horizon of the present time. By the successive inclusion of the laws of heredity and evolution the structure expressed by the phylogenetic (cladistic) system is superstructured by the contours of patristic and phenetic similarity, thus leading to their ideal reconstruction.

If we understand the “phylogenetic tree” as the universal ideal reproduction of life’s history on earth at the level of populations and species, then the phylogenetic (cladistic) system represents its basic structure, seen from the present time. As HENNIG (1950: p. 278) remarked, the phylogenetic tree of organisms—conceived as a diagram in which their phylogenetic (cladistic) relationships are related to their similarities in each conceivable direction—is a multidimensional structure which cannot be represented completely figuratively by a single picture. Therefore different projections of this polystructure are necessary for further cognition. These include

- the classification systems related to different strata of phenomena, to phenetic, patristic, and cladistic relationships;
- the classification systems related to different temporal horizons of the past as cross-sections across the phylogenetic tree;
- the representations of the phylogenetic tree which are expressed by the classification systems related to different levels of phenomena.

This totality is the conceptual network for catching the spatiotemporal diversity of the world of organisms. The phylogenetic (cladistic) classification system is the general reference system of the mental mastering of the organismic diversity *sub specie evolutionis*.

The factors, moving forces, and regularities of the historical development of life, of its evolution, are the subject-matter of the biological theory of evolution. The evolutionary theory is founded on the study of the present processes in order to explain the ideally reconstructed course of evolution. It answers the question how evolution works. Since Darwin’s *On the Origin of Species by Means of Natural Selection* (1859) it has passed through various stages of development. With Th. Dobzhansky’s *Genetics and the Origin of Species* (1937) began the stage of the Synthetic Theory of Evolution, of the synthesis of Darwinism and genetics.

Meanwhile there is some dissatisfaction with the Synthetic Theory and lively discussions are taking place. The Synthetic Theory became challenged for instance by the discovery of horizontal transmission of DNA-elements (“jumping genes”) and by the neutral theory of molecular evolution, by the theory of punctuated equilibrium, and by the evolutionary conceptions of RIEDL (1978) and of W.F. Gutmann and his group (GUTMANN and BONIK 1981). Apparently, the evolutionary theory is in a period of change—not its first and presumably not the last. As HULL (1978: pp. 338/339) remarked, there is not one set of propositions

(presented preferably in axiomatic form) which could be termed *the* theory of evolution. Instead there are several, incomplete, partially incompatible versions of the evolutionary theory currently extant. "I do not take this state of affairs to be unusual, especially in periods of rapid theoretical change. In general the myth that some one set of propositions exists which can be designated unequivocally as Newtonian theory, relativity theory, etc. is an artifact introduced by lack of attention to historical development and unconcern with the primary literature of science. The only place one can find *the* version of a theory is in a textbook written long after the theory has ceased being of any theoretical interest to scientists", he said.

ELDRIDGE (1982: pp. 77 and 78) sums up the debate in evolutionary biology as follows: "When all the dust settles from this latest episode of controversy in evolutionary theory, we will have a more accurate view of just how the evolutionary process works. That's the whole idea and what the game is all about. If evolutionary theory emerges in a somewhat altered form from the 'modern synthesis', some of us will feel victorious, and others will go to their graves in unyielding opposition. If the synthesis escapes unscathed, some of us will have tried in vain, but the theory will be all the stronger from its ability to withstand severe criticism . . . whatever emerges in the next ten years, it will be only a progress report."

The comments by Hull and by Eldredge demonstrate, in my opinion, very well the present general situation of the theory of evolution and the essence of the disagreements among evolutionists. These disagreements characterize the progress of the evolutionary thought in biology. It is an incomparably different situation from the beginning of our century, when KELLOG (1907: p. 9) resumed: "The fair truth is that the Darwinian selection theories, considered with regard to their claimed capacity to be an independently sufficient mechanical explanation of descent, stand to-day seriously discredited in the biological world. On the other hand, it is also fair truth to say that no replacing hypothesis or theory of species-forming has been offered by the opponents of selection which has met with any general or even considerable acceptance by naturalists. Mutations seem to be too few and far between; for orthogenesis we can discover no satisfactory mechanism; and the same is true for the Lamarckian theories of modification by the cumulation, through inheritance, of acquired or ontogenetic characters. *Kurz und gut*, we are immensely unsettled."

Now we need not be unsettled, but on the other side the Synthetic Theory in the shape it was presented in the textbooks of the fifties will

scarcely escape the present debate unscathed. To single out one point, there are problems with the factors of evolution. The term "factors of evolution" is used conventionally, without further reflection, as a common denotation of mutation, selection, recombination, isolation, fluctuations of populations, and genetic drift. Last but not least, molecular variability neutral to selection and jumping genes are not only reasons for enlarging the canon of the factors of evolution in the textbooks, but also for rethinking the concept of the factors of evolution on a theoretical level. This could be guided by the thoughts of ZAWADSKIJ and KOLČINSKIJ (1977). According to them, all aspects and components of the evolving life as well as the conditions and the moving forces of evolution may be comprehended as factors of evolution which can be singled out for a special study. The factors of evolution include every relatively discrete process and every feature of the organization of life which is part of the evolving substrate, or a cause or condition in the interactions that lead to the perpetual adaptive transformation of the populations.

According to ŠMAL'GAUZEN (1946), the totality of the factors of evolution can be divided into two groups. On the one hand there are the factors which make available and organize the substrate of evolution: the mutability, sexuality, individual variability, integration and isolation of populations and species and so on. These factors also include the horizontal transmission of genetic information and the variability at the molecular level; the neutral theory of molecular evolution is related to these factors. On the other side there are as opposite players those factors which act as moving forces, as the causes of evolution: struggle for life and natural selection. The relationships between these two groups of factors reveal the dialectical-contradictory character of evolution.

In the struggle for life, natural selection occurs which results from the interaction of many evolutionary factors and is the essential moving and directing force of evolution. Thus natural selection cannot be regarded as one of the several more or less distinct factors of evolution. It is fundamental for the understanding of natural selection that it not only favours or eliminates individuals, but realizes itself by means of its primary effect on the reproduction changes of individuals in the maintenance and advancement of populations and species. The modi of selection like stabilizing, directing, disruptive selection etc. always relate to populations, not to individuals. The present research and discussion has widened the Darwinian concept of environmental selection: the selection processes which influence the individual chances of survival and reproduction, occur at different internal organismic levels as well as superorganismic levels.

Examples of the latter are the kin- and group-selection of sociobiologists and the species selection of punctualists. What is selected is after all always individuals in populations.

A further aspect of the concept of the factors of evolution is the "evolution of evolution": the factors and the mechanism of evolution are not the same for all times and groups of organisms, but are themselves subject to evolution. As TOKIN (1935) already remarked, it is necessary to apply the historical point of view consistently to the process of evolution itself, and comprehend such concepts as (for instance) heredity and variability historically. Such a historical approach to the factors of evolution and their framework requires, as Zawadskij and Kolčinskij have demonstrated, a differentiation between the universal and the special factors of evolution. It should further be considered that the universal as well as the special factors of evolution can change quantitatively and qualitatively in different ways, and new factors can arise and old factors can disappear. To the universal factors of evolution belong for instance reproduction, mutability, struggle for life, and natural selection. Examples of special factors are the various modes of sexual and asexual reproduction, the symbiosis of the lichens or behaviour as pace maker in the evolution of animals. Moreover, in the case of the universal as well as the special factors of evolution, a distinction should be made between the main (essential or necessary and sufficient) factors which determine evolution (for instance mutability and natural selection), and additional factors, for instance, the tempo of the succession of generations, size and fluctuations of populations, fluctuation of environmental conditions etc.

On the basis of such distinctions between the factors of evolution the authors conclude that the theory of factors of evolution which until now was concentrated to more or less universal factors, should be developed further in two directions: (1) the special features of the factors and mechanisms of evolution of different great taxa like procaryotes, protozoa, or plants should be studied; and (2) the historical changes in the universal factors and the origin and disappearance of special factors in phylogeny should be studied. This approach to the factors of evolution includes also a changed understanding of the principle of actualism in the historical method: variaformism instead of uniformism.

Subtle anti-evolutionism attempts to throw doubt on the objective reality of evolution by disqualifying evolutionary biology by epistemological means. These attempts include the doctrine of the hypothetical past, the doctrine of ahistorical natural science versus lawless historical research, and the doctrine of the idea of descent as an aprioristic idea. The

doctrine of the hypothetical past is at home in empirical positivism and neopositivism. It states that the theory of descent is only a subjective hypothesis which is useful for ordering biological data but is impossible to prove or refute: statements concerning the past can as a matter of principle be nothing more than intellectual constructions since nobody has observed the events of the distant past. However, the question: "Did nature exist before the man?" is fatal to this doctrine (cp. LENIN 1977). In the seventies, on the occasion of the 100th anniversary of the geological glacial theory, the southern boundary of the Scandinavian inland-ice from the time of the quaternary formation was marked by 13 obelisks on the territory of the G.D.R. What is marked there? The boundary of the ice in the distant past or the boundary of an intellectual construction? The geologists wasted no thought to that question. Glacial theory and theory of descent have the same logic.

The doctrine of ahistorical natural science versus lawless historical research has its origin in the Freiburg school of Neo-Kantianism (W. Windelband, H. Rickert). It misses the relationships between the individual, particular, and universal, especially the relationship between individual historical events and the objective laws of development. Popper's anti-historicism and his confusing interpretation of Darwinism as a "metaphysical research programme" (POPPER 1974, RIEPPEL 1983) contain ingredients of both of the doctrines just mentioned.

The doctrine of the idea of descent as an aprioristic idea postulates that the theory of descent is only an interpretation of data to suit an aprioristic idea, an idea conceived prior to and independently of experience, which cannot be proved or refuted empirically *a posteriori* (MAY 1947). This doctrine is refuted by the history of biology which includes a long road of refutations of the conceptions of spontaneous generation. Redi's principle "omne vivum e vivo" is the quintessence of that long road. The concept of organismic evolution is a biological, a scientific, not a metaphysical concept.

References

- ELDRIDGE, N., 1982, *The Monkey Business* (Washington Square Press, New York).
GOULD, S.J., 1984, *Hen's Teeth and Horse's Toes* (Pelican Books, Harmondsworth).
GUTMANN, W.F. and BONIK, K., 1981, *Kritische Evolutionstheorie* (Gerstenberg, Hildesheim).
HENNIG, W., 1950, *Grundzüge einer Theorie der phylogenetischen Systematik* (Dt. Zentralverl., Berlin).

- HULL, D.L., 1978, *A Matter of Individuality*, Philosophy of Science 45, pp. 335–360.
- KELLOG, V.L., 1907, *Darwinism To-day* (Bell, London).
- KUHN, O., 1947, *Die Deszendenz-Theorie* (Meisenbach, Bamberg).
- LENIN, V.I., 1977, *Materialism and Empirico-Criticism* (Progress, Moscow).
- LÖTHER, R., 1972, *Die Beherrschung der Mannigfaltigkeit* (Fischer, Jena).
- LÖTHER, R., 1983, *Das Werden des Lebendigen* (Urania-Verl., Leipzig/Jena/Berlin).
- MAY, E., 1947, *Schöpfung und Entwicklung*, Zschr. f. philosoph. Forschg. 2, pp. 209–230.
- POPPER, K., 1974, *Darwinism as a Metaphysical Research Programme*, in: P.A. Schilpp, ed., *The Philosophy of Karl Popper*, Vol. 1 (Open Court, La Salle).
- PRÄGER, P., 1986, *25 Jahre Arbeitskreis Quartärgeologie der Gesellschaft für Geologische Wissenschaften der DDR — Ergebnisse und Aufgaben*, Mitteilg. Ges. für Geolog. Wiss. der DDR 14 (2,3), pp. 61–64.
- RIEDL, R., 1978, *Order in Living Organisms* (Wiley, Chichester/New York/Brisbane/Toronto).
- RIEPEL, O., 1983, *Kladismus oder die Legende vom Stammbaum* (Birkhäuser, Basel/Boston/Stuttgart).
- ŠMAL'GAUZEN, I.I., 1946, *Problemy darvinizma* (Sov. Nauka, Moscow).
- SNEATH, P.H.A. and SOKAL, R.R., 1963, *Principles of Numerical Taxonomy* (Freeman, San Francisco/London).
- TOKIN, B.P., 1935, *Vorwort*, in: *Probleme der Theoretischen Biologie* (INRA, Moscow/Leningrad).
- TSCHULOK, S., 1922, *Deszendenzlehre* (Fischer, Jena).
- ZAWADSKIJ, K.M. and KOLČINSKIJ, E.I., 1977, *Evoljucija evoljucii* (Nauka, Leningrad).

This Page Intentionally Left Blank

EVOLUTIONARY ALTRUISM AND PSYCHOLOGICAL EGOISM

ELLIOTT SOBER

Philosophy Department, University of Wisconsin, Madison, WI 53706, USA

1. Introduction

The idea that all human behavior is ultimately selfish has for a very long time had a considerable following in popular culture. For a very long time as well, philosophers have attempted to refute the idea by showing that it somehow rests on a confusion. Looking back on these two traditions, I find myself dissatisfied with both. I believe that psychological egoism is mistaken, but not for the reasons that the philosophical tradition has so far developed. So one problem I will address here is an issue in the philosophy of mind: how should we understand the thesis of psychological egoism?

A phenomenon more recent than the longstanding counterpoint concerning psychological egoism involves an issue in evolutionary theory. Evolutionists from Darwin down to the present have debated whether there are adaptations in nature that exist because they benefit the group. The main alternative position has been that traits evolve only because they benefit the organisms that possess them. These conflicting outlooks make different predictions about characteristics that benefit the group while placing the individuals who possess them at a disadvantage. Such characteristics have come to be called “altruistic”. If individuals compete only against other individuals, then altruism should not exist; if groups compete against other groups, then perhaps altruism will evolve. So the debate about the existence of altruism in evolutionary theory has focused on the plausibility of two ways of viewing the process of natural selection; the existence of altruism turns on the question of whether group selection has been real. In order to keep this issue separate from the debate about *psychological* egoism and altruism, let us call this question the problem of *evolutionary* altruism and selfishness.

There are a couple of obvious differences that should be noted that separate the psychological and evolutionary concepts of altruism. Individuals are psychological egoists or altruists by virtue of the kinds of motives they have. If I aim at hurting you but by accident do you some good, I am not thereby a psychological altruist. What counts in this concept is one's motives, not the effects of one's actions. On the other hand, the contrast between evolutionary altruism and selfishness involves the consequences of behavior; it is entirely irrelevant whether that behavior was proximally caused by a mind containing motives. In particular, the consequences that matter concern *fitness* — what matters is how an organism's behavior affects both its own prospects for survival and reproduction and those of the individuals with which it associates. Here again the evolutionary and psychological concepts part ways. If I give you a gift of contraceptives out of the goodness of my heart, this may show me to be a psychological altruist; however, in doing so, I may diminish your prospects for reproductive success, so my action may not be a case of evolutionary altruism.

Having separated these concepts, one cannot conclude that they have nothing to do with one another. After all, the mind is a cause of behavior; motives can produce actions that have consequences for survival and reproductive success. Perhaps the existence of psychological altruism in some way depends on how the human species evolved. If the human mind was shaped by a process of individual selection, then the mental adaptations that resulted came to exist because they helped individuals in their struggle for existence with one another. If psychological altruism implies a disadvantage to the individuals possessing that trait, does this mean that the trait should not exist, if our species evolved by a process of individual selection?

Several sociobiologists and evolutionists have taken this line. DAWKINS (1976: p. 3) rejects the idea of group selection and group adaptation and concludes that human beings are "born selfish". BARASH (1979: p. 167) likewise reasons from the primacy of individual selection to the impossibility of psychological altruism when he says that "real, honest-to-God altruism simply does not occur in nature . . . evolutionary biology is quite clear that 'What's in it for me?' is an ancient refrain for all life, and there is no reason to exclude *Homo sapiens*".

Here we see the idea that if all selection is individual selection, then psychological altruism is impossible. This conditional statement can be coupled with the assumption that its antecedent is true, in which case it follows that psychological altruism does not exist. Or the argument can be

run in reverse. If one believes the conditional and denies the consequent, the falsity of the antecedent follows.

This second pattern of reasoning—that psychological altruism is a reality, so it must be false that all selection is individual selection—is what we find in DARWIN'S (1872: pp. 163–166) views on human morality. Darwin's discussion in *The Descent of Man* begins by noting that psychological altruism involves a sacrifice in evolutionary self-interest:

It is extremely doubtful whether the offspring of the more sympathetic and benevolent parents, or of those which were the most faithful to their comrades, would be reared in greater number than the children of selfish and treacherous parents of the same tribe. He who was ready to sacrifice his life, as many a savage has been, rather than betray his comrades, would often leave no offspring to inherit his noble nature. The bravest men, who were always willing to come to the front in time of war, and who freely risked their lives for others would on average perish in larger numbers than other men.

Rather than concluding that psychological altruism does not exist, Darwin takes its existence as given and goes on to postulate an evolutionary mechanism that would account for it:

It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an advancement of well-endowed men will certainly give an immense advantage to one tribe over another.

When groups compete against groups, psychological altruism can evolve as a group adaptation. The trait exists because it is advantageous to the group, even though it is disadvantageous to the individuals possessing it.

There is no need to choose right now between Darwin's group selectionist explanation of the reality of psychological altruism and the individual selectionist argument that psychological altruism cannot exist. For the moment, I wish only to note a premise that both arguments assert: that psychological altruism cannot exist if all selection is individual selection. This is an assumption that bears looking at, once we have become clearer on what its component concepts involve.

In the next section, I will clarify the concept of evolutionary altruism. The idea is to bring clearly in view what the evolution of altruism requires; I will not take a stand on whether evolutionary altruism in fact exists in nature. I then will take up the concept of psychological egoism, again seeking to clarify what psychological egoism is. In conclusion, I will consider how evolutionary altruism and psychological altruism are related: does the one require the other, as the above two lines of reasoning

assume, or is it plausible to think that they have been decoupled in evolution? Here once more, the task is to clarify the different candidate explanatory relationships, rather than to reach some conclusion about the empirical facts of the matter.

2. Evolutionary altruism requires group selection

Evolutionary altruism is a comparative concept. A trait *A* is altruistic, relative to another trait *S*, which is selfish, precisely when the following conditions obtain: (i) within any group, *A* individuals are on average less fit than *S* individuals; (ii) groups of *A* individuals are fitter than groups of *S* individuals. For a characteristic to be altruistic (in the evolutionary sense) is for it to be disadvantageous to the individual who has it, but advantageous to the group in which it occurs.

These two defining facts about altruism are summarized in Fig. 1, which shows how an individual's fitness is influenced by whether it is altruistic or selfish, and also by whether it lives in a group in which altruism is common or rare. Note that the average fitness of individuals in a group (represented by the dotted line \bar{w}) is greater when *A* is common. Groups containing high proportions of altruists are more productive;

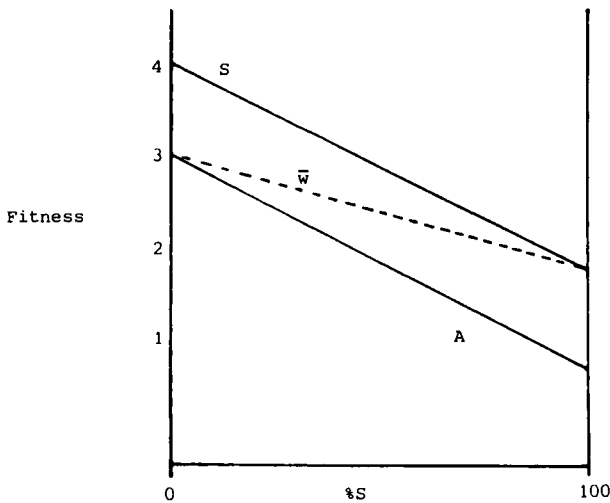


Fig. 1.

individuals in such groups have more babies *per capita* than individuals in groups in which altruism is rare. I assign fitness values ranging from 4 to 1 for convenience (these might be thought to represent the average number of offspring that individuals have when they live in groups of various kinds).

I said that evolutionary altruism and selfishness are “comparative” concepts to emphasize the following idea. Altruism implies the donation of some reproductive benefit, but not all donation counts as altruistic. Suppose a population contains individuals who donate one unit of benefit to the individuals with whom they live. Are these single unit donors altruistic? No answer is possible, until one says what other trait(s) the population contains. If the other individuals do not donate at all, then the single unit donors are altruistic. However, if the other individuals donate two units of benefit, then the single unit donors count as selfish. A trait is altruistic when it has certain fitness consequences *as compared with the other trait(s) found in the population*.

Another consequence of this definition of altruism is that not all cooperative behavior counts as altruistic. Consider a group of beavers who cooperate to build a dam. Does such cooperation count as altruistic? This is a question of how the cooperative behavior was related to other, alternative, traits that were present in the evolving population. This historical question may be difficult to answer, if all the beavers we now observe cooperate. The problem is a familiar one in investigating the workings of natural selection: natural selection requires variation, but often destroys the variation on which it operates. This means that the process has the unfortunate property of destroying some of the information needed to reconstruct its history.

Imagine that the ancestral population contained cooperators and free riders. A free rider is an individual who enjoys the benefits of a dam, but does not contribute to its construction or maintenance. If cooperation and free riding were present in the ancestral population, then cooperation would count as altruistic and free riding as selfish. Imagine instead that beavers who do not cooperate are severely penalized by the ones who build the dam. Here the choice is between cooperating and benefitting from the dam, and not cooperating and being severely penalized. If the penalty is severe enough, then there may be a quite selfish advantage in cooperating. In a given group, cooperators may be fitter than non-cooperators. If so, cooperation is a selfish — not an altruistic — trait.

This second scenario is the one that TRIVERS (1971) dubs “reciprocal altruism”. The present point is that reciprocal altruism is not altruism in

the sense defined above. Reciprocal altruism can evolve in a single population, but altruism cannot. Reciprocal altruists are *more* fit than non-reciprocators, and so reciprocity (cooperation) may reach 100% representation in the population. However, altruists, by definition, are *less* fit than selfish individuals in the same population, and so selection among individuals in a single population should lead altruism to disappear and selfishness to reach 100%.

In saying that reciprocal altruism is not altruism, I am not proposing some perverse and idiosyncratic redefinition of evolutionary concepts. Trivers himself notes that the point of his idea is “to take the altruism out of altruism”. Evolutionary altruism cannot evolve by individual selection alone; reciprocal altruism can. I emphasize the difference between them because they have very different implications about what the evolutionary process must have been like. By using the same terminology for both ideas, we run the risk of losing sight of the fact that two very different ideas are involved.

If altruism is defined in the way I have said, how could altruistic characters evolve? Within a single population containing both selfish and altruistic individuals, selfishness will displace altruism; and even if a population should by chance find itself containing only altruists, sooner or later a selfish mutant or migrant would appear and selfishness would be displaced. Within a single population, 100% altruism is not an *evolutionary stable strategy*, in the sense of MAYNARD SMITH (1982). In DAWKINS’ (1976) felicitous phrase, a population of altruists is vulnerable to “subversion from within”.

To see how altruism could evolve and be maintained, we need to consider not a single population, but an ensemble of them. Imagine that they vary in their local frequencies of altruism. Just to take an extreme case, imagine that there are two populations, one containing 99% altruists, the other containing 99% selfish individuals. Suppose that each group contains a hundred individuals. Below are the (approximate) fitness values of the individuals in each group and the fitness of each trait averaged across both groups; this information is simply read off from Fig. 1:

Group 1 (99% A)	Group 2 (99% S)
A: 3	A: 1
S: 4	S: 2
Global averages	
A: 3	
S: 2	

Note that within each group, altruists are less fit than selfish individuals. Yet, when one averages across the two groups, this inequality reverses: the average fitness of altruism is 3 (approximately) while the average fitness of selfishness is 2. This reversal of inequalities is a general phenomenon that statisticians have recognized in a variety of contexts; it is called Simpson's paradox and is absolutely central to the idea that group selection can permit altruism to evolve (SOBER 1984).

What will happen in the space of a single generation? Note that the system begins with altruism having a 50% representation in the two-population ensemble. We imagine that individuals reproduce and then die. To simplify matters we imagine that individuals reproduce uniparentally, and that offspring always exactly resemble their parents. Below are the numbers of individuals we would expect to find in the next generation:

Group 1 (99% A)	Group 2 (99% S)
A: 300	A: 1
S: 4	S: 200
Global census	
A: 301	
S: 204	

Note that altruism has increased from 50% representation in the global ensemble to something around 60%. Yet strangely enough, it is also true that altruism has declined in frequency within each group. In the first group it dropped from 99% to 98.7%; in the second it dropped from 1% to 0.5%. Simpson's paradox strikes again.

So much for the one generation calculation. What will happen if we follow the two groups through many generations? Before an answer is possible, we need to specify another assumption about how this system evolves, one having to do with whether (and, if so, how) groups send forth individuals to found colonies.

Let's imagine that the two groups hold together. They continue to exist as cohesive wholes; they do not fragment to found colonies. Since the individuals in both groups are mostly reproducing above replacement numbers, each group will increase in size. In each group selfishness is increasing in frequency, so sooner or later altruism must disappear from each group. This means that in the limit, altruism will be eliminated from the two population ensembles. The increase of altruism in the first generation from 50% to 60% was momentary. What goes up will come down, as subversion from within takes its toll.

Alternatively, imagine that once a group reaches a given census size, it fragments into a large number of small colonies. The members of an offspring colony are stipulated to all derive from the same parent population. Although I stipulated before that offspring *organisms* always exactly resemble their parents, I now imagine that the process of colony formation guarantees that an offspring *colony* may have a different frequency of altruism from that found in its parent. Colonies are formed by random sampling from the parental population, so sampling error will affect the composition of the offspring colonies.

What will happen now? A group with a high proportion of altruists will be more productive — will found more colonies — than a group with a low proportion of altruists. What is more, a parent group containing a high proportion of altruists will found colonies that display a variety of local frequencies of altruism. In fact, an altruistic parent colony will often have one or more offspring colonies that has a frequency of altruism that is higher than that possessed by the parent.

One more ingredient is needed for this group selection process to allow altruism to be stably maintained in the evolving system; it is time. It is essential that colony formation occur frequently enough, as compared with the rate at which selfishness displaces altruism within each group. To see the problem, suppose that selfish individuals are sufficiently fitter than altruists within any group, that a group that holds together for 25 organismic generations will see altruism disappear. If colony formation occurs more slowly than once every 25 organismic generations, it will come too late.

Thus altruism can evolve by group selection. It is essential that groups vary with respect to their local frequencies of altruism. What is more, it is important that similar organisms live together; note that in the simple two-population example, like lives with like. This is essential to allow Simpson's paradox to arise. Furthermore, it is important that groups found colonies. Without this, the enhanced productivity that arises from high concentrations of altruism will be for naught. Finally, it is important that groups found colonies fast enough, so that subversion from within cannot totally wipe out altruism.

When these conditions are satisfied, the evolving system will show the result of two processes that conflict with each other. Individual selection favors selfishness; group selection favors altruism. The result will be a compromise: neither altruism nor selfishness will be eliminated from the ensemble of populations.

All this is not to say whether group selection processes of the sort just

described have occurred, or whether they are responsible for many of the characteristics we observe in natural populations. That empirical issue is separate from the conceptual task of clarifying what evolutionary altruism is and showing how group selection makes possible what individual selection says is impossible.

3. Psychological egoism and the primacy of self-interest

I now shift from evolutionary theory to the philosophy of mind — from distal questions about how this or that trait arose in phylogeny to proximal questions about how preferences and motives produce actions in ontogeny.

Two truisms must be identified and set to one side. First there is the fact that in rational deliberation, agents choose that action which they believe will give them the most of what they want.¹ There is a trivial sense in which an altruist and a selfish individual both “do what they want to do”. This truism, I want to stress, does not show that psychological egoism is correct. The issue about egoism concerns *what* people want, not the trivial point that in rational deliberation, agents calculate on the basis of their own preferences (FEINBERG 1982).

The second truism is the fact that individuals almost always have preferences about what happens to individuals besides themselves. Altruistic individuals and those motivated by spite and malice have this in common (BUTLER 1726). In fact, quite selfish interactions with other people often involve such preferences. In ordinary buying and selling, the buyer wants to give money *to the seller* and the seller wants to transfer the goods *to the buyer*. Each has a preference about what should be true of the other, but this other-directedness does not mean that either agent is an altruist. The fact that we are other-directed (in this sense) does not establish that we are altruists any more than the fact that we rationally deliberate establishes that we are egoists.

To analyze the difference between psychological egoism and psychological altruism, I need to assume a distinction between self-directed and other-directed preferences. The former involves preferences about what happens to one's self; the latter concerns preferences about what happens

¹ I purposely make this description of rational deliberation vague so that it is common ground among a variety of more precise and competing theories. Here I am thinking of the debate between causal and evidential decision theories, on which see EELLS (1982).

to others. This distinction is not entirely unproblematic, since many preferences appear to be inherently relational. If I want to be a better volleyball player than you are, my preference seems to be neither entirely self-directed nor entirely other-directed. Rather, I simply want a certain relational fact to obtain.

I say “seems” and “apparently” because I think that this relational preference can be analyzed as an interaction between a self-directed and an other-directed preference. This does not mean that all relational preferences are so analyzable; in fact, I think that some are not. However, to estimate how restrictive my assumed distinction is, it is useful to see that it is not stopped short by the example before us.

My relational preference can be represented as a set of preferences concerning which of the following four situations I occupy. If my only concern is that I be better than you, then my ranking is as follows (higher numbers indicate better outcomes for the agent):

		Other	
		You are a good player	You are a bad player
Self	I am a good player	2	3
	I am a bad player	1	2

This preference ranking describes me as indifferent between the two outcomes shown on the main diagonal; I do not care whether we are both good or both bad.

This preference structure, I take it, accurately describes what it is for me to want the relational fact mentioned before to be true. Notice that I can now analyze this circumstance as an interaction between a self-directed and an other-directed preference. Whatever level of skill you display, I would rather be a good player than a bad one ($2 > 1$ and $3 > 2$); and whatever level of skill I possess, I would rather have you be a bad player than a good one ($3 > 2$ and $2 > 1$).² Therefore, my preference that a relational fact obtain between you and me is analyzable as an interaction between a self-directed preference and an other-directed preference.³

² Note that these conclusions could be reached, even if I divided the skill range into a larger number of finer categories; instead of “good” and “bad”, I might have described ten categories, ranging from “novice” to “expert”, or even a continuum.

³ If you think that being a good player is itself a comparative concept (good = better than average), the above point could be made by talking about wealth (= number of dollars). My desire to have more money than you is analyzable into a desire concerning my own level of wealth and a desire concerning yours.

I do not think that all relational preferences so neatly decompose into self-directed and other-directed components. Consider the preference I might have *to be different*. As far as volleyball goes, this might involve the following preference structure:

	You are a good player	You are a bad player
I am a good player	1	2
I am a bad player	2	1

What I want is to be bad if you are good and good if you are bad. This, I take it, is irreducibly relational. My assumed distinction between self-directed and other-directed preferences cannot handle this sort of example. I hope that the loss in generality is not too severe.

Besides distinguishing self- and other-directed preferences, I will assume that an agent's preferences can be described as a set of ranked outcomes displayed in tables like the ones just discussed. I will assume that rational agents choose actions in the following way. Each entry in a preference table corresponds to the overall merits of an action. Agents first decide which outcomes in a given preference table they can bring about. Such outcomes will be termed *available*. Among these, the agent selects the one that is most preferable. This is decision making under full information about the world; agents do not have to assign probabilities in describing what outcomes will obtain if they perform an action. If all the entries in the first table described above are available to me, I will choose the action that makes me a good volleyball player and you a bad one. The assumption that agents act with full information about the consequences of their actions restricts the generality of the model I will propose, but not in ways that matter to the points I want to make. The assumption just specified can be relaxed, without affecting the argument.

I want to begin by describing what it is for an agent to care not at all for the situation of others. Individuals of this sort I term *Sociopaths*:

		Other-directed preference	
		The other's situation is:	
		Good	Not-good
	One's situation is good	4	4
Self-directed preference	One's situation is not good	1	1

For such individuals, what happens to others *makes no difference*. Sociopaths faced with a decision problem in which all four outcomes are available will prefer acts that land them in the first row over ones that land them in the second; and they will be indifferent between actions that place them in the first column and ones that place them in the second.

The second sort of preference structure is the mirror image of the first. Individuals who care not at all for themselves, while wanting others to do well, I term *Kantian Robots*:

		Other-directed preference	
		The other's situation is:	
		Good	Not-good
	One's situation		
Self-directed preference	is good	4	1
	One's situation is not good	4	1

Although Sociopaths and Kantian Robots will sometimes behave quite differently, there is a decision problem in which they choose the same action. Suppose each is placed in a choice situation in which the only available outcomes are the ones that fall on the main diagonal. This means that only two actions are available; the first has the consequence that both self and other do well while the second results in both self and other doing badly. The Sociopath and the Kantian Robot will both prefer the first action, but for different reasons. Sociopaths choose an action that results in benefits flowing to others, but they do not perform this action *because* they care about others; Kantian Robots choose an action that provides a benefit to self, but not *because* they care about themselves.

I take it that both these preference structures are manifestly unrealistic descriptions of human motivation. Rarely are we totally indifferent to the interests of others; and rarely do we think of our own interests as counting for nothing. Each of these preference structures embodies *single factor* analyses of human motivation. If the problem of deciding whether people are altruists or egoists were simply the problem of deciding whether to think that people are Kantian Robots or Sociopaths, we would rightly reject both. It is true that Sociopaths are egoists of the most extreme sort; and Kantian Robots are extreme altruists. However, it is possible to describe altruism and egoism in such a way that neither presupposes a *single factor analysis of human motivation*.

An altruist may place weight on his or her own self-interest, but when self-interest and other-directed interests conflict, an altruist will prefer to

sacrifice self for other. I will call such individuals *Altruists* (with an initial capital). Kantian Robots and Altruists are both altruists (with lower case “a”). Thus, the Altruist’s preference structure is as follows:

		Other-directed preference	
		The other’s situation is:	
		Good	Not-good
Self-directed preference	One’s situation is good	4	2
	One’s situation is not good	3	1

The final preference structure is that of an Egoist who is not a Sociopath. I term such individuals Egoists even though they do not view the interests of others as counting for nothing. They are Egoists because their preference structure indicates that self-interest ought not to be sacrificed for the interests of others, should the two conflict:

		Other-directed preference	
		The other’s situation is:	
		Good	Not-good
Self-directed preference	One’s situation is good	4	3
	One’s situation is not good	2	1

All four preference structures lead to the same behavior when an agent is confronted with a main diagonal choice situation. However, they differ when the agent confronts an anti-diagonal choice situation. When the agent has to choose between benefitting self and benefitting other, Egoists and Sociopaths go one way while Altruists and Kantian Robots go the other. The first pair of preference structures differs from the second over whether self or other matters *more*. It is not that altruists (lower case “a”) only care about others and egoists (lower case “e”) only care about themselves. That contrast merely isolates what is special about Kantian Robots and Sociopaths. The categories differ as to which sort of preference matters *more* and which *less*; it is not necessary to think that one preference or the other matters not at all.

If this correctly identifies what distinguishes altruism from egoism, I think it is clear that some agents sometimes have altruistic preference structures. I am not saying that most of us always rate modest benefits to others as deserving more weight than gigantic benefits to ourselves. Few

of us would be willing to die to make someone smile. We should not think of the above preference structures as typifying agents' attitudes over all the choice situations they may confront. Each of us is a mixture of different preference structures; in different circumstances we accord different weights to self-interest and to the interests of others. However, the fact remains that most people are such that in some choice situations they have an altruistic preference structure. It is for this reason that psychological egoism is false as a generality about human behavior.

I have talked so far about preferences, not about the behaviors that those preferences engender. I think it is an open question how often people with altruistic preference structures produce altruistic actions. An altruistic action, I take it, is one in which the actor sacrifices his or her own interests for the sake of some other individual's interests. That is, an altruistic action is one that is produced in an anti-diagonal choice situation. It is not at all clear how frequently altruists find themselves facing problems of this kind.

Sometimes problems that superficially appear to involve such conflicts of interest in fact do not. Suppose I get to decide whether you or I receive some good — say, a cookie. My preferences count as altruistic if I would be willing to forego the good in order that you may have it. But suppose I would feel enormously guilty if I kept the cookie for myself and that I enjoy the smug glow of satisfaction by making the sacrifice for your sake. It may be true that letting you have the cookie provides more benefits to me than keeping the cookie for myself. In that case, foregoing the cookie and gaining pleasure coincide; I face a diagonal choice situation.

The point I want to make is that altruists may be so constituted that they rarely conceive of themselves as facing anti-diagonal choice situations. Perhaps they often would find it difficult to live with themselves if they kept the cookie for themselves. This does not conflict with the fact that they have altruistic preference structures. If they were forced to choose between receiving the guilt-free tickle of satisfaction without having the other benefit, or not receiving the sensation while the other benefits, they may sincerely prefer the latter. This makes them genuine altruists. However, the fact of the matter may be that the available options in the real world involve a strong correlation between benefits to others and benefits (including psychological benefits) to one's self. If so, altruists rarely get to act altruistically.

Altruism is a dispositional property, like "solubility". Altruists are disposed to sacrifice their own interests for the sake of others, should they be placed in an anti-diagonal choice situation; but it is a separate matter

how often they are placed in such choice situations. The parallel with solubility is this: soluble substances are disposed to dissolve when immersed; but it is a separate question how often such substances are in fact immersed.

Thus, my view is that it is much clearer that people sometimes have altruistic preference structures than that they act altruistically. This distinction does not mean that altruists suffer from chronic backsliding and weakness of the will. Rather, the problem is to obtain a correct understanding of how altruists conceptualize choice situations into available actions.

Just as the effects of altruistic preference structures are somewhat unclear, so too are their causes. To say that an individual has a given preference structure is to make a synchronic, not a diachronic, remark. There is an ontogenetic question here that I leave open. Perhaps the reason that adults have altruistic preference structures is that they were rewarded as children for helping others. It may be true that children begin life as egoists, but that the rewards they experience transform them into altruists. If so, one does not undermine the claim that adults are altruists by pointing out that they came to have the preference structures they do by being rewarded as children for acting in certain ways.

The point I now want to make is that what is true in ontogeny is also true in phylogeny. If it is a fact, as I claim it is, that individuals have altruistic preference structures, it is an open question how evolution could have produced that result. Darwin had one answer to that question; some sociobiologists seem to want to deny the phenomenon, because they think it is rendered impossible by a correct understanding of the evolutionary process. In the next section, I want to consider how psychological altruism is possible, even if the group selection scenario that Darwin considered does not correspond to the way natural selection works.

4. Psychological altruism without group selection

In the first section of this paper, I noted a conditional statement that both Darwin and some of his latter day followers have implicitly endorsed in their discussion of psychological altruism. It is the idea that if a trait *T* is found in a population, the explanation must be that there was selection for the presence of *T* ancestrally. Darwin reasoned that since psychologi-

cal altruism is a reality, there must have been selection for its presence. Since individual selection would not make psychological altruism advantageous, he sketched an account in which group selection does the trick. Barash, on the other hand, starts with an individual selectionist point of view about evolution and reasons that there cannot be selection for psychological altruism. From this he concludes that the trait cannot be a reality.

There is a third option that deserves a hearing. It is the idea that psychological altruism is spin-off, a "spandrel", in the language of GOULD and LEWONTIN'S (1979) influential paper. Maybe psychological altruism exists because it was correlated with other characters which themselves were selectively advantageous to the individuals possessing them. Conceived of in this way, psychological altruism exists because of individual selection, even though there was no selection for it.⁴

Let us consider an example of how this proposal might be fleshed out. Psychological altruism is probably too broad and heterogeneous a category to be treated as a single characteristic. Let us consider a somewhat narrower example of how altruistic feelings sometimes function. Here I have in mind the fact that human beings sometimes adopt unrelated children and raise them as their own. Sometimes the adoptive parents cannot be biological parents; sometimes they can, but choose to adopt nonetheless. It is pretty clear that there is no selfish reproductive interest that is served by this practice. From a kin selection point of view, it would make more sense for prospective adopting parents to help their nieces and nephews. Adoptive parents behave in ways that do not make sense from the perspective of a theory that says that each and every behavior of an organism must be fitness maximizing.

That there was no selective advantage in wanting to adopt unrelated children does not make the presence of that trait utterly mysterious from an evolutionary point of view. Perhaps there was an evolutionary advantage in having individuals feel the sentiments we call maternal and paternal; and perhaps a spin-off correlate of this trait is the inclination to adopt children in certain circumstances.

To make this more precise, consider the fitness values we would plausibly assign to the following four combinations of psychological characteristics:

⁴ For discussion of the difference between the concepts of *selection for* and *selection of*, see SOBER (1984).

	Want to adopt in some circumstances	Do not want to adopt
Care about one's biological offspring	<i>a</i>	<i>b</i>
Do not care about one's biological offspring	<i>c</i>	<i>d</i>

I take it to be uncontroversial that $a > c$ and $b > d$. Regardless of one's feelings about adoption, there will be a selective advantage in caring about one's own biological offspring, should one have any. Now let us imagine that the desire to adopt is never fitness enhancing: $a \not> b$ and $c \not> d$. Does this mean that the wish to adopt cannot emerge in a population subject to individual selection?

This does not follow. Suppose that there is a correlation between being inclined to care about one's own biological offspring (should one have any) and being inclined to want to adopt, should certain circumstances arise. If the correlation is perfect, then all the individuals in the population will have both sentiments or neither; each individual will have a fitness value of a or of d . If $a > d$, then the two traits will evolve together, the one being fitness enhancing, the other being neutral or even slightly deleterious.

This simple pattern of argument shows how psychological altruism could evolve without the need for a group selection hypothesis. Even if there is no individual advantage in being a psychological altruist, the preference structure that goes by that name may have been correlated with other traits that represented an individual advantage.

It is arguable, I suppose, that the desire to adopt was phenotypically "silent" until rather recently. Perhaps human beings only recently encountered circumstances of the sort that would lead them to want to adopt. If so, the presence or absence of the inclination to want to adopt would have made no selective difference during the earlier period in which we are imagining these sentiments to have evolved. When the environment changed in the historically recent past, the inclination to want to adopt suddenly expressed itself in behavior. Note here that a trait coding for a given behavior can evolve by individual selection during a given time period, even though it has no effect on the behaviors of organisms. Again the key idea that makes this possible is correlation.

Other spin-off explanations can be invented. SINGER (1981) conjectures

that the altruism encoded in human moralities is a spin-off from the more general faculty of human reason. Perhaps the ability to reason abstractly—to consider and be moved by rational considerations—evolved by individual selection. Once in place, this ability may have many spin-off consequences which confer no adaptive advantages. Perhaps the ability to compose fugues and the ability to solve differential equations are examples. More to the point, perhaps a susceptibility to being moved by rational considerations in moral deliberation is a spin-off effect of a more general ability to reason. If so, individuals will find themselves moved to accord weight to the interests of others, not just to their own interests, because rationality indicates that they ought to. This form of psychological altruism—embodied in the willingness to act on impersonal principles—would then exist because of individual selection, but not because there was direct selection for being a psychological altruist.

5. Concluding remarks

Evolutionary altruism is a historical concept. If a trait is an example of evolutionary altruism, this implies something about how it could have come into existence. In particular, the implication is that it could not have evolved by being favored under individual selection, but might have evolved by group selection. The concept of psychological altruism has no such implication. That trait is understood in terms of a given preference structure; this structure has implications about how an individual will behave in various choice situations; but it leaves open what the proper explanation is of the fact that people have such preference structures.

In separating evolutionary and psychological altruism in this way, I am not saying that the one has no explanatory relationship to the other. My claim is that such a connection needs to be argued for as an empirical thesis. Even if one thinks that evolutionary altruism is impossible because one holds that all selection is individual selection, it does not follow that psychological altruism is impossible. The idea of evolutionary spin-off is meant to illustrate how psychological altruism could emerge in a species of organisms whose evolution is governed by individual selection.

As in so many areas of evolutionary investigation, a very serious problem is posed by how the phenotype of an organism is to be segmented into “characters”. “Altruism” is a category of common sense, one which applies in a multiplicity of choice situations. There is no *a*

priori reason why the dispositions lumped together by common sense under this rubric should form a single evolutionary unit. Perhaps different aspects of this phenomenon had quite separate evolutionary histories. If so, we misconceive the evolutionary problem by demanding a single explanation of the existence of "altruism".

One way in which this common sense category should be segmented into its separate aspects is already visible in the discussion of psychological altruism provided here. It is quite unsatisfactory to describe an individual as an altruist, full stop. What is true is that an agent has an altruistic preference structure in a given choice situation. I may be prepared to sacrifice my interest for yours in one context where the stakes are modest, but not in another, where they are higher. We should not ask whether and why people are altruistic, but whether and why they are altruistic in some choice situations but not in others.

It would not be surprising if a great many of these more specific questions turn out to have only a trivial connection with the facts of our evolution. Specific altruistic dispositions may have their origins in culture and custom, not in the changes in gene frequencies that evolutionary theory seeks to explain. When we ask why some individuals in some circumstances are altruistically disposed, while other individuals in other circumstances are not, the answer may be that there are cultural differences between the individuals and situations in the first case and those in the second. It is conceivable, I grant, that natural selection may have shaped the genetic characteristics shared by these individuals in such a way that people are genetically determined to respond differently to the two situations; and it also is conceivable that the first individuals differ genetically from the second, in ways determined by natural selection, and this genetic difference explains why they behave differently. It is also possible that evolutionary considerations do not explain the difference at all.

If this "non-evolutionary" possibility is the right one, it still will be true that human evolution has made it possible for people to behave differently in the two situations. This is not to say that the facts of evolution explain why people behave differently. There is yet another truism that needs to be recognized here: everything that human beings do is consistent with the facts of human evolution. This does not mean that the facts of human evolution explain everything that human beings do. The facts of evolution may show why *X* and *Y* are possible behaviors while *Z* is not; but this will not explain why some people do *X* while other people do *Y*.

These last remarks must remain speculative, since we do not at present

understand very well how to segment that amorphous collection of inclinations we term “altruistic” in a principled way. It remains to be seen whether the phenomena of psychological altruism can be consolidated by a univocal treatment within theories of ontogeny (i.e. in psychology) and within theories of phylogeny (i.e. in evolutionary theory).

References

- BARASH, D., 1979, *The Whisperings Within* (Penguin, London).
- BUTLER, J., 1965, *Sermons*, in: L. Selby-Bigge, ed., *British Moralists*, vol. 1 (Dover Books, New York).
- DARWIN, C., 1871, *The Descent of Man, and Selection in Relation to Sex* (J. Murray, London).
- DAWKINS, R., 1976, *The Selfish Gene* (Oxford University Press, Oxford).
- EELLS, E., 1982, *Rational Decision and Causality* (Cambridge University Press, Cambridge, UK).
- FEINBERG, J., 1982, *Psychological egoism*, in: J. Feinberg, ed., *Reason and Responsibility* (Wadsworth, Belmont, CA).
- GOULD, S. and LEWONTIN, R., 1979, *The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme*, *Proc. R. Soc. London* 205, pp. 581–598.
- MAYNARD SMITH, J., 1982, *Evolution and the Theory of Games* (Cambridge University Press, Cambridge, UK).
- SINGER, P., 1981, *The Expanding Circle* (Farrar, Straus and Giroux, New York).
- SOBER, E. 1984, *The Nature of Selection* (MIT Press, Cambridge, MA).
- TRIVERS, R., 1971, *The evolution of reciprocal altruism*, *Q. Rev. Biol.* 46, pp. 35–57.

10
Foundations of Psychology
and Cognitive Sciences

This Page Intentionally Left Blank

VISION AND MIND

VADIM D. GLEZER

*Pavlov Institute of Physiology, Academy of Sciences of the USSR, Nab. Macarova 6,
199034 Leningrad, USSR*

The main idea of my paper is as follows. I'll try to discuss in terms of neural organization how the brain creates the basic categories by means of which it comprehends the world.

The traditional approach to semantic problems is through language. The majority of investigators agree that thought and speech are formed by deep semantic structures which are determined by the construction of the brain. However, the semantic models are based only on linguistic material. The analysis of such models shows the insufficiency of the linguistic approach based only on the method of the black box. I will discuss the functional organization of the nervous mechanisms owing to which some universals corresponding to basic categories of thinking and speech are formed on the basis of sensory raw-material. The visual system is suitable for such an approach because vision has been relatively well investigated. The second reason is that vision is the basic supplier of sensory information in man. However, the main reason is as follows. There is a lot of evidence that vision is the basis of thinking. The experimental data will be discussed in my paper. But there are many indirect indications. For example, it is well known that the difficulty of introducing the new ideas in physics at the beginning of this century was caused by the impossibility of visualizing new concepts. The impossibility to create the visual image "particle-wave" resulted in the rejection of new ideas during the formation of relativistic physics. This fact shows that the visual brain underlies our thinking at least on the common sense level.

But perhaps the strongest evidence in favour of this opinion is that we can say "I see" when we mean "I understand".

The investigations of vision give us a lot for the understanding of higher psychical functions. There is a widely accepted view that vision and other sensory modalities serve only to introduce information to the brain which

uses this information. The detector theory favours this point of view to a great extent. The detectors segmentize the signal, analyze it and detect in it various components.

The alternative point of view is that vision is also thinking but concrete, objectual thinking. In visual perception the visual world is segmentized into objects. The objects and the relationships between the objects (that is, spatial relationships) are described. The description of the objects is invariant with respect to different transformations of the object on the one hand, and concrete—that is, the full description—on the other hand. Thus the philosophical categories of the abstract and the concrete are already present in visual thinking.

We can show now, to some extent, the neurophysiological mechanisms underlying such an organization.

The concept of the modules of neocortex has recently gained wide recognition. The area of the neocortex is subdivided into a mosaic of quasi-discrete spatial units. These spatial units are the modules which form the basic anatomical elements in the functional design of the neocortex. There are approximately 3 million modules in the human neocortex.

A module is a group of neurons having certain functional and morphological unity. The module may be described as a vertically oriented group of neurons with strong vertical connections and weak horizontal ones. According to the concept of modules, the neocortex is formed by a mosaic of uniform iterative units. This concept is based mostly on morphological investigations. Mountcastle says that the data on the cytoarchitectonic and external connections of modules is evidence that the module may be regarded as an objective mechanism of conscious perception.

Eccles states that the modules are the neural correlates of conscious experience and of mental events.

These ideas seem to be logical and true, but they do not give a formalized model of the functioning of modules. What operations of processing information do the modules perform? Using vision, I'll try to show that the information processing performed by the modules really underlies the basic cognitive processes.

The modules of the visual cortex have been examined best of all. The Nobel prize winners Hubel and Wiesel have shown that the cells of the visual cortex have a very peculiar organization. The cell responds only to a line of a certain width and orientation. The reasons for this are clearly seen if we consider the organization of the receptive field of the cells. The

zones of RF have an elongated form. The central zone excites the cell, the peripheral zones inhibit the cell when light falls on the RF. Therefore if a light bar is placed along the central zone, the cell responds. If the line is placed orthogonally to the optimal orientation, the cell does not respond. The response of the cell is equal to the integral of the product of the weighting function of the cell and the distribution of light in its RF. A module contains cells, the RFs of which have different orientation and width. The responses of the cells of the modules give the description of the image.

On the basis of our data we have proposed an alternative organization of the module. We have shown experimentally that the narrow-width elements of the module are united and form a grating pattern. This small difference leads to important consequences for understanding the role of the module as a device for processing information. It may be assumed that the cells of the module perform the piecewise Fourier-expansion of the image.

The organization of such a module can be shown schematically as follows. The RFs of the cells of the module overlap an area of the field of vision. The cells are tuned to different orientations and different spatial frequencies. Each cell of the module computes the coefficient of Fourier-expansion. If the device performs the Fourier-description it must contain several harmonics. It means that the weighting functions of the cells must be formed by 1, 2, 3 and more cycles. The experiments show that weighting function of the cell comprises several cycles. We have never seen more than four cycles in a weighting function.

The same result can be achieved by other data. At the same time, the data reveal some new aspects of the problem.

If the cells of the module perform the Fourier-description of the image, then the interval of expansion must be a constant value. It means in our case that the RFs of the module must have a constant size, though it is well known that the sizes of the RFs of visual cortex vary to a great extent. We can assume that the module is comprised of RFs of the same size but there are modules of different sizes. This means that the distribution of the sizes of RFs must be discrete. The experiment supports this assumption: the RF-sizes at eccentricity from 0 to 6° cluster at one half octave intervals and form a discrete distribution. It means that the modules of all sizes coexist in the same areas of the visual field. The results are statistically highly significant.

Let us compare these facts with the distribution of the spatial frequencies to which the cells are tuned. The optimal frequencies also

cluster at one half octave intervals and form a discrete distribution. The distribution is formed by linear cells which can perform the Fourier-analysis. It is interesting to compare it with the distribution of optimal spatial frequencies of non-linear cells obtained by the American investigators Pollen and Rönner. The distributions practically coincide. The non-linear cells can calculate the power spectrum. I'll discuss it later.

Let us perform a simple operation. If we divide the value of the size by the value of the period of optimal frequency we will get the number of periods in the weighting function. The averaged values characterize the number of periods in the weighting functions of the cells forming a module *irrespective of the size* of the module. We have got the series: 1, 2, 3, 4. So for every size of the module the number of the harmonics is limited to four. To be exact, we have an additional term equal to 1.41, or a root-square of two. This term was observed mostly with diagonal orientations of the RFs in accordance with the predictions of the two-dimensional piecewise Fourier-expansion.

Let us summarize these facts and propose the following hypothesis. The visual field is overlapped by nets of modules of different sizes. The module of an appropriate size is selected for every image or subimage in the field of vision. Subimage is a part of an image. The subimage in a complex image may be an image by itself. The segmentation of the visual field is a complex process. An essential role in it is played by the non-linear cells of the module. These cells calculate the piecewise power spectrum. The images or subimages differ by textures which have different local power spectra. The non-linear cells can extract the figure from the background using this property as they measure the local power spectrum. There are also some other mechanisms for segmentation of the visual field and selection of an appropriate module for the image. These mechanisms are: directional cells, binocular cells, color cells.

When the module for an object or a part of an object is selected, the linear cells of the module give its full description in terms of Fourier-coefficients.

Such a model of the visual cortex has interesting properties. It creates premises for the uniform invariant description of an object, on the one hand, and for the description of its spatial parameters, on the other hand. The invariants are the spectral coefficients of the module irrespectively of its size and position. The spatial characteristics (that is, size and position) are defined by the number (address) of the module.

It is interesting to note that the object is described only by 4 harmonics.

It is a very poor description. But it was found in psychophysical experiments that 4 harmonics are enough for recognizing an image. The experimental data allow us to assume that when we want to recognize fine details in the image we use a module of a smaller size. For example, face is recognized by the information produced by a given module, but the form of the nose is evaluated with the aid of a module of a smaller size.

The reality of such an organization has been supported by many psychophysical and behavioural experiments conducted in our laboratory.

The premises which are created by the module organization are used by the two basic mechanisms of the visual system. In experiments performed on monkeys and cats it has been shown that the mechanism of invariant description is localized in the inferotemporal cortex of the monkey and in the 21 field and the dorsolateral and ventrolateral regions of suprasylvian sulcus in the cat. This mechanism, according to our assumptions, uses the output of the module. The mechanism of the description of spatial relationships is localized in parietal cortex in field 7. This mechanism, according to our assumptions, uses the number of the module. Let us discuss briefly one experiment as an example.

The monkey was taught to differentiate one image from other images. Normal or transformed images were shown. In the intact animal the more the image is transformed — decreased or increased in size — the lesser is the percentage of right reactions. It is a natural result if we keep in mind that the visual system of the monkey does not know what we want. The two basic mechanisms are opposed. The mechanism of invariant description states that the transformed image is the same as the image by which the animal was trained. The mechanisms of the description of spatial relations states that it is quite a different image. The more the image is transformed, the lesser is the percent of responses according to which the transformed image is the same as the initial image.

After extirpation of occipito-parietal cortex in one group of monkeys the animals chose the image regardless of its transformation. The mechanism of concretization does not exist any longer. Only the mechanism of invariant description remains.

A directly opposite result was obtained when the inferotemporal cortex was extirpated. The animals respond only to the initial image and do not recognize it when it is transformed.

It was shown in other experiments that in this case the animal does not see the image, but its behaviour is governed by the spatial features which were memorized in the parietal cortex.

The other evidence was obtained in psychophysical experiments. A very good correlation between the proximity of the images in psychophysical and spectral spaces was found if the properties of the module were taken into account. The most important property for centering the image in a module is as follows. The weighting functions of cells in a module have a 90° phase off set. It means the existence of sinusoidal and cosinusoidal functions. If the module is centered relative to the image so that the cosinusoidal low harmonics give the minimal response, the correlation is very high.

In one experiment the matrices of mixing up images for two real subjects and for the model were compared. In the model the distance between two images was calculated as Euclidean distance between the Fourier-transformations of the images. The beginning of the coordinates was chosen so as to minimize the phase spectrum of the low frequencies. The correlations between the elements of the matrix of mixing up and the model distances were very high -0.94 for one subject and -0.96 for the other subject. The correlations are highly statistically significant.

Of course it is a great oversimplification to state that the categories of abstract and concrete on the visual level are based only on the dichotomy — parietal-temporal cortices. The second dichotomy also exists — left-right hemispheres. In many psychophysical experiments performed in our laboratory it has been shown that the methods of visual information processing in the left and right hemispheres are very different. We can summarize these facts in the following way. The left hemisphere describes the image by the discriminant method — with the aid of discriminant features. The left hemisphere uses the invariant properties of a module's organization. We assume that the modules converge on the device localized in the left inferotemporal cortex. The cells of the inferotemporal cortex capable of learning, from a hyperspace in which hyperplanes or discriminant features extract a volume corresponding to a certain image. The set of the features gives an invariant description of the object.

On the contrary, the right hemisphere does not make the invariant description. The right hemisphere uses the structural method. It means that the direct outputs of modules which correspond to the descriptions of subimages or images are united in the right temporal cortex with the aid of spatial operators of the right parietal cortex in concrete images.

I repeat that the problem of the interrelationship of the dichotomies: parietal-temporal and left-right, is a very complicated one. It deserves a separate discussion. Nevertheless this problem only complements and

develops the general construction but does not contradict our main conclusion about the role of modules and the two basic mechanisms.

Let us sum up our conclusions. The modules segmentize the visual field into separate subimages and images. The outputs of modules are their descriptions in a form which allows the following abstract and concrete description. The numbers or addresses of the modules are used for the indication of spatial characteristics of the object, such as size and position, or for describing the spatial relations between the objects. The last operation is performed by the operators of the occipito-parietal cortex. The analogue of such an operator in the theory of artificial intellect is a frame—a term introduced by Minsky. We have shown in behavioural experiments that some areas in the occipito-parietal cortex are responsible for describing spatial relations. The animal with extirpation of one of these areas cannot distinguish between large and small objects. With another lesion it cannot distinguish between the situations where the objects were in exchanged positions.

In the right hemisphere the operators describe the relations between the parts of an image, in the left—between the images creating scenes.

On the basis of all these mechanisms the visual brain creates a model of visual world. The information kept in the model is highly ordered, which facilitates the process of extracting the information from the model and comparing it with the information coming from the eye. This process may be termed visual thinking.

The mutual functioning of the two basic mechanisms produces both a generalized abstract description and a concrete description of an object. Every act of visual perception includes the comparison of new information with a well-ordered model of the world kept in the visual brain, and it means simultaneously the completion and development of the model. We can conclude therefore that the act of visual perception must be treated as an act of objectual non-verbal thinking.

Now let us discuss how these mechanisms are being developed and adapted for logical thinking.

Assuming that the categories elaborated on the sensory level are deep universals, lying behind the categories of mind and language, we can approach the understanding of the organization of higher psychical functions in a formal way.

There is enough evidence in favour of this concept. Let us discuss at first very briefly the mechanism of classification. After the extirpation of inferotemporal cortex in monkeys, the mechanism of invariant classification is damaged. It means that the mechanism of invariant description is

damaged when the images are projectively transformed. The lesion of an anterior part of the inferotemporal cortex leads to the impossibility of higher abstractions. The monkey cannot unite different images in one class. The animal cannot understand that different images are followed by identical reaction. This experiment is a model of the inability to unite a capital and a small letter (a and A). This process is already a non-visual classification but yet non-verbal.

Let us discuss now the mechanism of relations. It is plausible to suggest that the spatial operators and logical operators should be identical in organization. For example, the following statements are very similar. The 1st statement: the objects A and B are at the same height. The 2nd statement: the objects A and B are the same. In both cases the operator compares the responses of two modules or, to be more exact, the descriptions of the two objects received from classification mechanism.

This assumption is fully supported by clinical data, described by Head and Luria. The occipito-parietal lesions in the left hemisphere lead to the impossibility of understanding both the spatial relations between the objects and the complex logical-grammatical constructions serving for the coordination of the details as a whole. The patients with parietal lesions recognize the objects but do not conceive the spatial relations between them. At the same time, they do not understand temporal relations (spring is before summer), comparative relations (Katja is more blonde than Sonja), transitive acts (lend money to somebody), or logical relations (brother of the father).

Evidently in all these cases the device named "frame" by Minsky is absent. A frame is a structure of data representing a stereotype situation. The same structure may be used for different situations depending on the data filling the cells of the frame. If the cells of the frame are filled up with images, the frames constitute an extra- or intra-personal space. If they are filled up with concepts, the frames form a conceptual space. There is some experimental evidence to suggest that the cells of the frame are filled up with the aid of the mechanism of selective attention. From this point of view the mechanism of selective attention in the visual process must be treated as setting in motion with the aid of feedback the local operators measuring the spatial relations. The address may be directed on very different levels of the visual system, thus embracing different areas of the visual field or different spectral components of image description and so on.

Let us sum up what has been said about the postparietal cortex. The PP cortex forms an extra- and intra-personal space on the basis of frames,

and it is a regulator of selective attention. The latter is analogous to setting in motion the operators describing spatial relations both between the elements of the image and between the images. Thus this mechanism gives a concrete expression to the situation.

But the cells of a frame may be filled up with any nominations — with abstractions of a much higher order than images. The frame with the same organization achieves a qualitatively higher sense. For example, Minsky has shown that the description of the spatial situation in the room is essentially the same as the story told by a boy how he had bought a present for a friend for his birthday.

Thus the understanding or comprehending is analogous to the revealing of a part of the model of the world performed by unfolding the frame, that is, by filling up its cells.

The process of recognition consists not only in classification or comparison with a pattern kept in memory. When we look at a branch of a tree we do not only recognize it as a branch. We see the concrete branch with all its peculiarities. We can see how buds, shoots and leaves are situated. It means that we understand or comprehend quite a new situation. Both of the basic mechanisms serve this purpose.

It is interesting to note that the scientists investigating the theoretical problems of brain functioning, for example, those who investigate the problem of artificial intelligence often oppose the two approaches: global, holistic and local, atomic. For example, Minsky states that global ideas such as microworlds or problem spaces are diversions from the traditional atomic approach favoured both by behaviourists and by those who are oriented to mathematical logic and try to describe knowledge as the conglomerate of simple elements.

The structure of the model described here includes both approaches and relates them to different material structures, showing the underlying neurophysiological mechanisms.

It is interesting to compare these results with some linguistic conceptions. Roman Jakobson showed that the occipital lesions lead to the damage of the paradigmatic function of the language, and the frontal lesions lead to the damage of the syntagmatic function.

As to the paradigmatic function, we can directly compare object lexicon and the visual images. As was shown in our experiments, an ordered system of invariant visual images exists. The multi-dimensional space of images is divided with the aid of a hierarchy of complex invariant features. The object lexicon is analogously organized with the aid of paradigmatic relations and hierarchical oppositions. The oppositions in

visual images may be compared to oppositions of the expressions in a vocabulary. For example, such hierarchical oppositions exist: animate-inanimate, running-flying and so on. As was discussed before, the brain localization in this case is the same or very nearly the same.

The syntagmatic function of the brain may be divided into the predicative syntagmatic and the syntagmatic of object nominations. I will not discuss the predicative syntagmatic function which is based on the mechanisms of movements localized in the anterior part of the brain. The mechanisms of the syntagmatic of object nominations as mechanisms of the connection of nominations in congruous judgements are localized in the occipito-parietal cortex. As was shown before, in these cases, patients with such lesions understand neither spatial relations nor logical-grammatical constructions. For example, the patient says: "I understand the meanings of "brother" and "father", but I do not see the difference between "brother of the father" and "father of the brother"." It is interesting to discuss the problem together with other categories of speech such as grammar or vocabulary. At parietal lesions spatial and logical relations are damaged. In speech, the constructions connected with prepositions and cases are damaged. It is evident that all the prepositions reflect spatial relations. Analogously, cases reflect the relations between objects in concrete thinking. Different forms are used in different languages. In English relations are expressed by prepositions and by the genitive, in Russian—both prepositions and cases, in Finnish—only cases, in Bulgarian—only prepositions, in Turkish—by particles introduced into a word, in Chinese—by the sequence of words in a sentence. However, regardless of the form in different languages, it is the same universal-spatial relations in concrete thinking and logical relations in abstract thinking. The grammatical categories may also be traced in the second mechanism. The paradigmatic function is reflected variously in different languages. The opposition of nominations in a vocabulary is one of grammatical categories. For example, the opposition of "animate-inanimate" is expressed in grammars of different languages by endings, articles, pronouns.

Thus the grammar of a language is a reflection and manifestation of deep structures and is based on the two basic mechanisms which were defined in the investigation of the visual system. The universals elaborated by these mechanisms are the result of the construction of the brain developed in evolution. If the construction of brain were different as a result of some other process of evolution, then the model of the world would be different and the deep semantic structures would be based on

some other universals. For example, if instead of a system of piecewise Fourier-analysis, serving for extracting and describing objects, the global Fourier-expansion of the whole visual field were used, then we would think not in terms of single objects and concepts and relations between them but in terms of whole scenes.

It is interesting to note that in some semantic models based on linguistic material it is stated that a thought has no grammatical structure. The neurophysiological analysis leads to a different conclusion.

Let us summarize all that has been said before. The mechanisms of thinking require a well-ordered storage of information in the nervous system, allowing a fast retrieval of the requisite codes and the operation with them. The orderliness is achieved thanks to the two basic forms of the language organization—vocabulary and grammar are embedded in the construction of the brain, and in the first place—of the visual brain. Grammar is organized differently in different languages, but as a set of rules for holding information in a sensory model of the world it is based on the two main mechanisms of the sensory brain.

References

- CHOMSKY, N., 1968, *Language and Mind* (New York, Toronto).
ECCLES, J.G., 1981, *The modular operation of the cerebral neocortex considered as the material basis of mental events*. *Neuroscience* 6(10), pp. 1839–1856.
GLEZER, V.D., 1985, *Vision and Thought* (Leningrad) (In Russian).
LURIA, A.R., 1962, *The Basic Problems of Neurolinguistics* (Moscow) (In Russian).
MINSKY, M., 1975, *The Psychology of Computer Vision* (New York).
MOUNTCASTLE, V.B., 1978, *The Mindful Brain* (The MIT Press).

This Page Intentionally Left Blank

11
Foundations of
Social Sciences

This Page Intentionally Left Blank

RATIONALITY AND SOCIAL NORMS

JON ELSTER

University of Chicago, Illinois, U.S.A., and Institute for Social Research, Oslo, Norway

A persistent cleavage in the social sciences opposes two models of man conveniently associated with Adam Smith and Emile Durkheim, *homo economicus* and *homo sociologicus*. Of these, the former is supposed to be guided by instrumental rationality, while the behaviour of the latter is dictated by social norms. The former is “pulled” by the prospect of future rewards, whereas the latter is “pushed” from behind by quasi-inertial forces.¹ The former is easily caricatured as a self-contained, asocial atom, and the latter as the mindless plaything of social forces.

In this paper I try to clarify the concept of rationality and that of social norms. I spend more time on the latter, since there is much less agreement on its definition. On the analysis I propose, there are situations in which rationality and social norms prescribe different courses of action. In some of these, people adhere to the canons of rationality. In others, rationality yields to social norms. My task here is not to state the conditions under which the one or the other can be expected to dominate. To do so would require the elaboration of a substantive social theory, going well beyond methodological concerns. My main purpose is simply to state the distinction and argue for its empirical relevance. In particular, I shall have to defend it against various reductionist attempts to explain social norms as being in a more fundamental sense rational, optimal or adaptive.

1. Rational behaviour

My analysis of the concept of rationality is guided by its use later in this paper as a contrast to the idea of social norms. Rational agents act

¹ For a useful exploration of this contrast, see the analysis of educational choices in Italy in GAMBETTA (1987).

consistently and efficiently, always searching for the best means to achieve their ends. They are sensitive to variations in the environment, such as changing relative prices of goods. They are aware of efforts by other agents to achieve goals that may affect their own, and try to shape their own behaviour so as to gain maximally or suffer minimally from these other actions. In all of these respects, they can differ from agents guided by social norms.

The theory of rational action from which these statements are derived is a familiar one.² Rationality is defined as a relation between action, belief, desire and evidence. An observed action is rational if it is the best means to realize the agent's desire, given his beliefs about relevant factual matters. It is, in a word, optimal. It is, moreover, performed *because* it is believed to be optimal, so that merely accidentally optimal actions do not count as rational. Furthermore, the beliefs of the agent are themselves subject to a rationality constraint: they must be well grounded in the evidence available to the agent. The amount of evidence, finally, must also be scrutinized from the point of view of rationality or optimality. The definition of a rational action, in fact, stipulates optimality at three different levels.³ The action must be optimal relative to the given beliefs and desires. The beliefs must be optimal relative to the evidence. The amount of evidence collected must be optimal relative to the agent's desires and his other beliefs.

From this it follows that there are two main ways in which rationality can fail, and leave a wedge for other principles of action.⁴ The theory may fail to yield unique prescriptions and hence predictions. Also, people may fail to conform to its prescriptions. In one phrase: the concept of rationality may be indeterminate, or people may be irrational.

The first kind of failure can arise at each of the three levels of optimality. Given his beliefs and desires, the agent may face several courses of action that are equally and maximally good. More radically, if his preferences are incomplete there may be *no* optimal action. Given his evidence, he may be unable to form the relevant factual beliefs. In particular, he may not be able to form reliable beliefs about the expected

² For a fuller analysis, see ELSTER (1986).

³ In addition, one might want to require that the agent's desires be themselves optimal. The concept of optimal or rational desires remains elusive, however; much more so than the other levels of optimality. For a brief discussion, see Ch.I of ELSTER (1983).

⁴ For a fuller discussion, see ELSTER (forthcoming).

behaviour of other agents whose choices influence his own outcome. Finally, he may not have the knowledge that would allow him to assess how much evidence to collect.

The second kind of failure can also arise at each level. At the first, weakness of will may prevent people from choosing what they believe is the best means to realize their ends. At the second and third, cognitive inefficiencies and affective disturbances such as wishful thinking may interfere with belief formation and evidence gathering.

The theory of social norms, or any other theory of human action such as Herbert Simon's theory of satisficing behaviour, might arise to fill either of these gaps. Perhaps people rely on social norms when and to the extent that rationality is indeterminate.⁵ Or perhaps social norms are the culprit when people behave irrationally. On the first hypothesis, the theory of rationality and the theory of social norms supplement each other, with the latter clearly subservient to the former. On the second hypothesis, which will be the focus of the present paper, the theories offer competing explanations of human behaviour. The competition is local rather than global. The issue is not which of the two theories gives the right account of human behaviour in general, but whether any particular piece of behaviour is best accounted for by the one or the other theory.

2. Social norms

The canons of rationality tell people what to do if they want to achieve a given end. Social norms tell people what to do or not to do, either unconditionally or conditionally upon other people's behaviour. Rationality is consequentialist: it is concerned with outcomes. Social norms are non-consequentialist: they are concerned directly with actions, for their own sake. This characterization does not preclude that norm-guided actions may have good consequences, nor that norms are sustained by the consequences of respecting and violating them, through positive and negative sanctions by other people. It does preclude, however, fine-tuned sensitivity to outcomes. The theory of social norms

⁵ See MARDEN (1986) for a discussion of the view that "customary forces operate within a 'range of indeterminacy' left by the action of market forces" (p. 137).

predicts, for instance, that behaviour should not change when the feasible set of actions expands.⁶

To lay the groundwork for the analysis I begin by offering some examples of norm-guided behaviour.

(1) Simple, paradigmatic cases of social norms are those regulating manners of dress, manners of table and the like.⁷ They are regulated by two apparently contradictory principles: to be like other people, and to differentiate oneself from other people. Hence they define *status groups* in Max Weber's sense.

(2) The following is a plausible, although apparently irrational pattern of behaviour. I would not be willing to mow my neighbour's lawn for twenty dollars, but nor would I be willing to pay my neighbour's son more than eight dollars to mow my own, identical lawn. This has been explained by the psychological difference between opportunity costs (income foregone) and out-of-pocket expenses.⁸ The explanation might also, however, be the presence of a social norm against mowing other people's lawns (and performing similar services) for money.⁹

(3) Norms against incest, cannibalism and other "acts contrary to nature" exist in most, but not all societies.¹⁰ Cannibalism is nevertheless allowed under certain specific circumstances, and is then regulated by social norms. The custom of the sea, for instance, has been quite clear: "What sailors did when they ran out of food was to draw lots and eat someone."¹¹

(4) The norm of voting, of doing one's civic duty in national elections, is widespread in democratic societies. Without it, it is doubtful whether there would be any voluntary voting at all.¹²

(5) In many societies there is a social norm against living off other people which explains why workers sometimes refuse wage subsidies, although they may accept and indeed lobby for much more expensive subsidies to their firms.¹³

⁶ BECKER (1962) notes that the theory of social norms is powerless if the feasible set *contracts* so as to exclude the behaviour prescribed by the norm.

⁷ For numerous examples, see BOURDIEU (1980).

⁸ THALER (1980).

⁹ I owe this suggestion to AMOS TVERSKY.

¹⁰ EDGERTON (1985) has a full discussion of these and similar examples.

¹¹ SIMPSON (1984: p. 140).

¹² BARRY (1979: Ch. 2).

¹³ For examples see ELSTER (1988).

(6) There is a social norm against offering other people money to get their place in a queue, even when both parties would gain by the transaction. This is but one of many examples of social norms against using money to buy what money is not supposed to be able to buy.¹⁴

(7) A "code of honour" among workers may prevent an employer from hiring new workers at lower wages, for instance because the old workers refuse to engage in the required on-the-job training of the new even at no costs to themselves. As a result, unemployment may ensue.¹⁵

(8) A more destructive code of honour is that which requires people to engage in vendettas, if necessary over generations, and often with gruesome results.¹⁶

(9) Some societies have had norms of strict liability for harm, regardless of intent or mitigating circumstances. "For example, should a man's wife die in childbirth, the husband was liable for her death; had he not impregnated her, the Jalé said, she would not have died."¹⁷

(10) A more general norm is the widespread "norm of reciprocity" which enjoins us to return favours done to us by others.¹⁸ The potlatch system among the American Indians is a well-known instance. According to one, contested, interpretation the potlatch was something of a poisoned gift: "The property received by a man in a potlatch was no free and wanton gift. He was not at liberty to refuse it, even though accepting it obligated him to make a return at another potlatch not only of the original amount but of twice as much."¹⁹

(11) There are important social norms that tell people to cooperate in situations that offer a strong temptation to take a free ride on the cooperation of others. The norm may be unconditional, or conditional upon the cooperation of others. Unconditional cooperation is embodied in the norm of "everyday Kantianism": cooperate if and only if it is better for all if everybody cooperates than if nobody does so. The norm forbids one to consider the (actual) consequences of one's own behaviour, and enjoins one instead to consider the (hypothetical) consequences of a set of actions. Conditional cooperation corresponds to a norm of fairness:

¹⁴ For other examples, see WALZER (1983).

¹⁵ AKERLOF (1980). See also SOLOW (1980).

¹⁶ For a description of a famous case, see RICE (1982). I am grateful to Robert Frank for drawing my attention to this example.

¹⁷ EDGERTON (1985: p. 161).

¹⁸ GOULDNER (1960).

¹⁹ Helen Codere, cited after PIDDOCKE (1965).

cooperate if and only if all or most others cooperate. This is a norm against free riding, which is conditional but not consequentialist. The action is not made conditional upon consequences, but on the behaviour of other people. Since these norms are substantively important and very central in recent debates over norms, I discuss them separately later.

I now review two objections to the view that social norms affect behaviour in a way that conflicts with outcome-oriented rationality. First, one might deny that social norms affect behaviour. Secondly, one might argue that when people obey social norms they are in fact concerned with outcomes.

A fundamental problem that arises in the analysis of social norms is to what extent they have real, independent efficacy and to what extent they are merely rationalizations of self-interest.²⁰ Is it true, as argued by early generations of anthropologists and sociologists, that norms are sovereign and people little more than their vehicles? Or is it true, as argued by more recent generations, that rules and norms are nothing but raw material for strategic manipulation?²¹ Let me note at the outset that whatever judgment one finally reaches on this issue, the appeal to norms cannot be *merely* strategic. "Unless rules were considered important and were taken seriously and followed, it would make no sense to manipulate them for personal benefit. If many people did not believe that rules were legitimate and compelling, how could anyone use these rules for personal advantage?"²² We must also, however, reject the "over-socialized view of man" and the "over-integrated image of society" that underlies the writings of Durkheim or Parsons.²³ The truth must be somewhere between these extremes, leaving scope for disagreement about which of them is closest to the truth.

The miserable Ik of Uganda provide a good illustration both of the

²⁰ A third position is advocated by CANCIAN (1975). She argues that norms as defined by Parsons and others have no relation whatsoever to behaviour, neither as *ex ante* generators of action nor as *ex post* justifications of action. Among her subjects in a Maya community, she found no correlation between norm clusters elicited by comparison questions and choices in three sets of alternative actions: (1) whether an individual farmed nearby or took advantage of the new road and farmed far away; (2) whether he sent his children to school; and (3) whether he used Western doctors in addition to native curers. She also provides a subtle and thoughtful discussion of an alternative conception of norms, with emphasis on norms as rules for validation of one's social identity by others.

²¹ For a brief history and clear statement of this distinction, see EDGERTON (1985: Ch. 1).

²² EDGERTON (1985: Ch. 1, p. 3).

²³ WRONG (1961).

reality of norms and of their manipulability. Discussing the institutions of gift and sacrifice among the Ik, Colin Turnbull writes that

These are not expressions of the foolish belief that altruism is both possible and desirable: they are weapons, sharp and aggressive, which can be put to diverse uses. But the purpose for which the gift is designed can be thwarted by the non-acceptance of it, and much Icier ingenuity goes into thwarting the would-be thwarter. The object, of course, is to build up a whole series of obligations so that in times of crisis you have a number of debts you can recall, and with luck one of them may be repaid. To this end, in the circumstances of Ik life, considerable sacrifice would be justified, to the very limits of the minimal survival level. But a sacrifice that can be rejected is useless, and so you have the odd phenomenon of these otherwise singularly self-interested people going out of their way to "help" each other. In point of fact they are helping themselves and their help may very well be resented in the extreme, but it is done in such a way that it cannot be refused, for it has already been given. Someone, quite unasked, may hoe another's field in his absence, or rebuild his stockade, or join in the building of a house that could easily be done by the man and his wife alone. At one time I have seen so many men thatching a roof that the whole roof was in serious danger of collapsing, and the protests of the owner were of no avail. The work done was a debt incurred. It was another good reason for being wary of one's neighbors. [One particular individual] always made himself unpopular by accepting such help and by paying for it on the spot with food (which the cunning old fox knew they could not resist), which immediately negated the debt.²⁴

This kind of jockeying for position is widespread. There is, for instance, a general norm that whoever first proposes that something be done has a special responsibility for making sure that it is carried out. This can prevent the proposal from ever being made, even if all would benefit from it. A couple may share the desire to have a child and yet neither may want to be the first to lance the idea, fearing that he or she would then get special child-caring responsibility.²⁵ The member of a seminar who suggests a possible topic for discussion is often saddled with the task of introducing it. The person in a courtship who first proposes a date is at a disadvantage.²⁶ The fine art of inducing others to make the first move, and of resisting such inducements, provides instances of instrumentally rational exploitation of a social norm. The norm must have solid foundations in individual psychology, or else there would be nothing to exploit. The crucial feature of these norms is their conditional character. *If A does X, there is a norm that B do Y.* An example is the norm of reciprocity. Hence, if *Y* is burdensome *B* has an incentive to prevent *A* from doing *X*. *If A does X, there is a norm that he also do Y.* An

²⁴ TURNBULL (1971: p. 146).

²⁵ I am indebted to Ottar Brox for this example.

²⁶ WALLER (1937).

example is the norm of special responsibility for the proposer. Hence, *A* has an incentive to abstain from doing *X*.

Social psychologists have studied norms of equity, fairness and justice to see whether there is any correlation between who subscribes to a norm and who benefits from it. Some findings point to the existence of a "norm of modesty": high achievers prefer the norm of absolute equality of rewards, whereas low achievers prefer the norm of equity (i.e. reward proportionally to achievement).²⁷ More widespread, however, are findings which suggest that people prefer the distributive norms which favour them.²⁸ This corresponds to a pattern frequently observed in wage negotiations. Low-income groups invoke a norm of equality, whereas high-income groups advocate pay according to productivity.

More generally, there is a plethora of norms that can be used to justify or limit wage claims. To justify wage increases, workers may refer to the earning power of the firm, the wage level in other firms or occupations, the percentage wage increase in other firms or occupations and the absolute wage increase in other firms or occupations. When changes are being compared, the reference year can be chosen to make one's claim look as good as possible. Conversely, of course, employers may use similar comparisons to resist claims for wage increases. Each of these comparisons can be supported by an appropriate norm of fair wages. There is a norm of fair division of the surplus between capital and labour, a norm of equal pay regardless of the type of work, a norm of equal pay for equal work, a norm of preservation of status (or wage differences), a norm of payment according to productivity etc. Workers will tend to invoke the norms which justify the largest claims, and employers those which justify moderate raises. Yet, once again, these references cannot be purely and merely opportunistic, for then they would carry no weight with the adversary party. Also, each party is somewhat constrained by the need to be consistent over time in the appeal to norms.

When I say that manipulation of social norms presupposes that they have some kind of grip on the mind since otherwise there would be nothing to manipulate, I am not suggesting that society is made up of two sorts of people: those who believe in the norms and those who manipulate the believers. Rather, I believe that most norms are shared by most people — manipulators as well as manipulated. Indeed, efficacious mani-

²⁷ G. Mikula, *Studies in Reward Allocation*, cited in SELTEN (1978). See also KAHN *et al.* (1977).

²⁸ DEUTSCH (1985: Ch. 11); MESSICK and SENTIS (1983).

pulation usually requires belief in the norm you are invoking. Purely cynical use of norms is easily seen through: "man merkt die Absicht und wird verstimmt."²⁹ Rather than manipulation in this direct sense, we are dealing here with an amalgam of deception and self-deception. At any given time we believe in many different norms, which may have contradictory implications for the situation at hand. A norm that happens to coincide with narrowly defined self-interest easily acquires special salience. If there is no norm handy to rationalize self-interest, or if I have invoked a different norm in the recent past, I may have to act against my self-interest. My self-image as someone who is bound by the norms of society does not allow me to pick and choose indiscriminately from the large menu of norms to justify my actions, since I have to justify them to myself no less than to others. At the very least, norms are soft constraints on action.³⁰ Sometimes they are much more than that, as shown by the case of vendettas.

The second objection concerned the non-consequentialist nature of social norms. Since deviating from a norm can have unpleasant consequences, in the form of negative social sanctions, could we not say that one follows the norm to bring about the absence of these sanctions?³¹ Also, sometimes conformity with norms elicits positive approval, and not simply the absence of disapproval. Moreover, application of the norms — e.g. expressing approval or disapproval — is also an instrumentally useful act. Whenever there is a first-order norm enjoining or forbidding some action, there is a meta-norm enjoining others to punish defectors from the first-order norm.³² Failure to express disapproval invites disapproval and hence makes it individually rational to sanction deviants.

If everyone else accepts the norm and is willing to act on it, by conforming to it or by punishing non-conformers, it may indeed be individually prudent to follow the norm and to apply it to others. But we can still sustain the distinction between rational, outcome-oriented behaviour and behaviour guided by social norms, by comparing norms with the mutually shared expectations that characterize much of strategic behaviour. In games with a unique equilibrium point and full informa-

²⁹ This is a main theme in VEYNE (1976) and, following him, in Ch. II of ELSTER (1983).

³⁰ For a similar argument see FØLLESDAL (1981).

³¹ I am not here considering internalized norms, which can sustain behavior in the absence of any external sanctions. This aspect of norms is briefly considered in the final section of the paper.

³² For a discussion of metanorms, see AXELROD (1986).

tion, everybody will conform to everybody else's expectations. *A* expects *B* to do *X* and expects *B* to expect *A* to do *Y* because if they do neither has an incentive to act otherwise. The behaviour is not guided by social norms, because the expectations can be derived endogenously from the assumption of rational actors. By contrast, if *A* exacts revenge for the murder of his brother because he knows that his family expects him to, the expectations are exogenous and given prior to the interaction. In strategic action we deduce expectations about what people will do from their preferences for outcomes. With other outcomes, expectations may differ. In norm-guided action, expectations are simply given and outcomes have no role to play. In strategic action, for instance, threats must be credible in terms of outcomes. Norm-guided threats, by contrast, are *ipso facto* credible. Sometimes, as in the cases studied by T.S. Schelling, the credibility of a rigid code of behaviour is instrumentally useful. When the code is shared, however, the outcome may be mutually disastrous.

Strategic behaviour constitutes one useful contrast to norm-guided behaviour. Action guided by social norms may also be contrasted with action guided by "private law".³³ Consider the heavy smoker who tries to quit, but constantly finds himself backsliding. One way out of his predicament is to construct unbreakable rules for himself. Rather than limiting his intake of tobacco to, say, three cigarettes a day, he decides to become a total abstainer, living by William James's advice "Never suffer a single exception". This principle can provide a bright line, a focal point that allows for no manipulation or ambiguity, and hence is invulnerable to the numerous self-deceptive tactics at our disposal. The behaviour is rigid, inflexible, insensitive to circumstances and to some extent to outcomes. From an outcome-oriented point of view he would be better off if he smoked an occasional cigarette, but this consideration has no force.

"Private law" is to some extent analogous to social norms. The behaviour is not outcome-oriented, but guided by the idea that certain acts are inherently bad. One recoils from them in horror, without pausing to ask whether and when the reaction is justified. The difference is that the private law is an individual construct, not the result of socialization. The Freudian superego has elements of both. Many people have unbreakable rules against smoking, gambling or drinking because they have been

³³ I borrow this phrase from GEORGE AINSLIE (1982, 1984, 1986), to whom I am heavily indebted in what follows.

brought up to think that these activities are inherently bad even in small doses. Others have been led to construct such rules for themselves, as they found that moderation did not work for them. In practice, the two cases can be hard to distinguish from each other.³⁴

3. Norms of cooperation

Norms of cooperation in collective action situations are an especially important subclass of social norms. I assume familiarity with the general structure of collective action problems, which are closely related to the n -person Prisoner's Dilemma. Since my argument turns upon features of non-standard cases, some preliminary remarks are nevertheless in order.

I define a collective action problem by the following features. First, there are a number of individuals each of whom has the choice between cooperating and not cooperating, e.g. not polluting and polluting. Secondly, there is no external authority to impose negative or positive sanctions. Thirdly, cooperation is costly for those who engage in it. Specifically, the costs to the individual cooperator are always larger than the benefit he derives from his contribution. Here, "benefit" must be taken in a narrow, self-interested sense, which does not take account of the impact on other people. Fourthly, the benefits of cooperation depend solely on the number of cooperators. Finally, there exists a level of cooperation—i.e. a number of cooperators—that makes everybody better off than if nobody cooperated.

These propositions are fully compatible with the following non-standard possibilities. First, at some levels of cooperation the costs of cooperation to the cooperator may exceed not only his benefits, but the sum of all the incremental benefits accruing to group members. Typically, this tends to happen at very low and very high levels of cooperation. Secondly, individual acts of cooperation may have a negative net impact not only on the cooperator, but also on other people, both at low and at high levels of cooperation. The cooperative act may be directly harmful

³⁴ Cf. the following passage from LEVY (1973: pp. 184–85): "[An] expressed motive for the involvement in religion is 'protection from one's own impulses to bad behavior'. Teiva, for example, says that all villagers are religious (although not enough to save themselves from hell) because they remember the savage pagan behavior of their ancestors, the wars and cannibalism (matters which missionary teachings constantly emphasized when they portrayed the salvation from savagery brought by religion), and being afraid of backsliding, use religion to protect themselves from doing evil."

to others, or it may deprive them of other, more valuable services which the cooperator would otherwise have provided. I give some examples later.

The starting point for my discussion is the following claim. If people always cheated and defected when it was in their rational self-interest to do so, civilization as we know it would not exist. Conversely, the existence of civilized society points to the presence of restraining forces. Social norms against cheating and defection could be one such force, but I shall also consider other possibilities.

First, however, I must argue for the proposition that rational self-interest would induce cheating in many contexts. Descartes argued that the requirements of *long-term* self-interest coincide with those of morality. People who help others tend to receive assistance in return, and as a consequence are better off in the long run than are people who exploit each and every opportunity for cheating. The cause of cheating, in this view, is neither rationality nor self-interest, but myopia and weakness of will. The remedy for cheating is self-control or "private law" rather than social norms. As mentioned, these two mechanisms are closely related, but I want to insist on the analytical distinction between them. There is, for instance, no social norm to engage in life-extending forms of physical exercise although it is clearly in the long-term self-interest of the individual to do so (unless it is true, as I read somewhere, that for each hour of exercise life expectancy goes up by about fifty minutes).

The Cartesian idea has been made more precise by the theory of iterated games elaborated over the last few decades.³⁵ A central conclusion is that under certain combinations of parameter values, cooperation conditionally upon cooperation by others in the previous round is an equilibrium point of the iterated game. We cannot infer, however, that a cooperative equilibrium will in fact be realized by rational actors. For several reasons—multiple equilibria, incomplete information, wrong parameter values—cheating may prevail. Nor can we infer that when people do cooperate under the conditions specified by the theory, they do so because it is in their rational self-interest. Long-term self-interest may be pre-empted by morality or by social norms. In my view, there are neither good theoretical nor good empirical reasons for believing that people often cooperate out of rational self-interest.

This view does not, however, yield the conclusion that cooperation is due to social norms. People might have rational, non-selfish motivations

³⁵ See notably AXELROD (1984) and TAYLOR (1987).

for cooperating with others. A utilitarian motivation to maximize total welfare would ensure cooperation in many situations, as would concern for close relatives and friends. These are outcome-oriented motivations, that lend themselves to ends-means calculations of the usual kind. A utilitarian would not cooperate, for instance, if the expected number of other cooperators is so small or so large that the cost to him of cooperating exceeds the sum of the incremental benefits to others. Consider call-in campaigns to support public television. If many people call in, the impact of each additional call is reduced, whereas the cost to the caller goes up because of the time he has to wait before getting through.

The norms of cooperation—everyday Kantianism and the norm of fairness—differ in several respects from moral motivations. Most importantly, they are not outcome-oriented. This follows at once from the statement of the norms, but the point is worth making at greater length. Acting on either of these two norms can have bad consequences for average welfare or even for everybody's welfare.

Consider first everyday Kantianism. In one sense this norm refers to outcomes, viz. to the state of the world that would be brought about if everyone acted in a certain way. It does not, however, refer to the outcome of the individual act of cooperation or non-cooperation. It does not allow consideration of the external circumstances, such as the expected number of other cooperators, that determine whether an individual act will in fact have good or bad consequences. Because it neglects these circumstances, everyday Kantianism can lead to bad outcomes in several kinds of cases. A population of Kantians might suffer costs of cooperation that are not justified by the benefits created. More importantly, benefits might actually be destroyed, if universal cooperation leads to people tripping over each other's feet, thus reducing the efficiency of cooperation. Individual Kantians can also do harm to others. Unilateral disarmament can increase the risk of war, if it creates a power vacuum which other states rush in to fill. Unilateral acts of heroism or sacrifice can give authorities or employers an excuse to crack down on non-participants as well as participants.

Consider next the norm of fairness. On the one hand, this norm allows non-cooperation even when the benefits to others would largely exceed the cost to the individual. On the other hand, it can prescribe cooperation when each additional contribution reduces average welfare, and even when it reduces everybody's welfare. Joining the army in wartime is an example. Those who stay home to work in vital industries may feel that

they are violating the norm of fairness. If all who want to join were allowed to join, the war effort as a whole might suffer. A group of friends who are cleaning up after a party might actually get the job done faster if some of them were to relax instead with a drink, but the norm against free riding might overwhelm considerations of efficiency.

4. Varieties of reductionism

I have been concerned to argue for a *prima facie* difference between rational behaviour and action guided by social norms. Against this one might conceive of various reductionist strategies. Of these, the simplest is just to deny that norms are ever the proximate causes of action. Instead, norms are seen as strategic tools in *ex post* rationalization of self-interest. I have already discussed and rejected this view, arguing that it is internally inconsistent. In addition, there is much direct evidence for norm-guided behaviour. Studies of voting³⁶ and of tax evasion³⁷ provide examples of uncoerced, uncoordinated cooperative behaviour that is best explained by assuming the operation of norms of cooperation. Countless anthropological studies offer hard evidence for the existence and efficiency of social norms.³⁸

More complex forms of reductionism would grant that social norms may be the proximate cause of behaviour, but argue that the norms can themselves be explained by some form of optimality reasoning. I shall distinguish between three reductionist strategies of this variety, in terms of collective rationality, individual rationality and inclusive genetic fitness. I believe that the strategies fail, singly and jointly. No single reductionist strategy can account for all social norms. Some social norms cannot be reduced by any of the three strategies. Actually, I shall be arguing for somewhat weaker conclusions, to the effect that we have no good reasons to believe that the strategies can succeed, singly or jointly.

The crudest form of reductionism appeals to collective optimality, either in the sense of Pareto-optimality or in the sense of maximizing total welfare. Social norms are to be explained by the fact that everyone, or at least the average individual, benefits from them. To state this view, I can do no better than quote a passage by Kenneth Arrow:

³⁶ BARRY (1979).

³⁷ LAURIN (1986).

³⁸ EDGERTON (1985) offers many examples and a thoughtful discussion focused on the issue of the efficacy of norms.

It is a mistake to limit collective action to state action . . . I want to [call] attention to a less visible form of social action: norms of social behavior, including ethical and moral codes. I suggest as one possible interpretation that they are reactions of society to compensate for market failure. It is useful for individuals to have some trust in each other's word. In the absence of trust, it would become very costly to arrange for alternative sanctions and guarantees, and many opportunities for mutually beneficial cooperation would have to be foregone . . .

It is difficult to conceive of buying trust in any direct way . . . indeed, there seems to be some inconsistency in the very concept. Non-market action might take the form of a mutual agreement. But the arrangement of these agreements and especially their continued extension to new individuals entering the social fabric can be costly. As an alternative, society may proceed by internalization of these norms to the achievement of the desired agreement on an unconscious level.³⁹

Although second to none in my admiration for Arrow's work, I cannot help finding this passage astonishingly naive. Endowing "society" with a capacity for self-regulation which operates through the individual unconscious to set up norms that compensate for market failures, would seem to be a blatant violation of methodological individualism. Perhaps the passage should not be taken too literally. Yet it is easy to think of writers who would subscribe to the literal interpretation. Functionalist, cybernetic, system-theoretic and Marxist conceptions of society share the idea that society is an organic entity with laws of self-regulation and self-development of its own.

There are three objections to this account of norms. First, not all market failures are "solved" by appropriate norms. Many societies would have greatly benefited from a norm against having many children, and yet such norms are virtually non-existent. Norms against corruption are sorely lacking in many societies. Examples could be multiplied.

Secondly, many social norms are not collectively optimal.⁴⁰ A striking example is the maxim in Jewish law that when a life-saving good cannot be given to everyone, it should not be given to anyone.⁴¹ The norm against buying places in the queue would similarly seem to be a pointless prohibition of potential Pareto-improvements. Norms allowing or prescribing suicide or vengeance are not in any obvious sense collectively

³⁹ ARROW (1971: p. 22). For other examples see ULLMANN-MARGALIT (1977: p. 60) and NORTH (1981: Ch. 5).

⁴⁰ The force of this objection might be weakened by an argument that currently suboptimal social norms may have been introduced because and at a time when they were optimal, and owe their continued existence to an inertial lag (ARROW 1971: p. 22, NORTH 1981: p. 49). In some cases this suggestion may well be correct, but there is no reason to think that present vices are always the legacy of past virtues.

⁴¹ ROSNER (1986: pp. 347-348).

optimal. The small Italian village studied by Edward Banfield was characterized not only by the absence of norms against corruption, but by the presence of norms against public-spirited behaviour.⁴² Again, examples could be multiplied.

Thirdly, the fact that a social norm is collectively optimal does not by itself provide an explanation of its presence. An explanation requires the demonstration of a *mechanism* whereby the collectively beneficial effects tend to maintain the norms and the behaviour that caused them. Consider the norm against early marriages found in many societies. It is tempting to explain this norm by the collectively beneficial effects of having smaller families. But it could also, more simply, be a norm against premature gratification, on a par with norms against drinking or gambling. It might be rational for the individual to postpone marriage, to afford having a family.

For a more complex example, consider the norms of cooperation. Although action according to these norms may occasionally have bad consequences, their overall effect is surely beneficial. Yet it is far from clear that they owe their presence to these effects. It may be more plausible to seek the explanation in individual psychology.

Everyday Kantianism may owe more to a psychological mechanism that has been called “everyday Calvinism”.⁴³ This is the confusion of causal and diagnostic efficacy, or the fallacious belief that by acting on the symptoms one can also change the cause. If a predeterminist doctrine like Calvinism could lead to entrepreneurship, it can only have been via the magical idea that manipulating the signs of salvation could enhance the certainty that one was among the elect. Similarly, it has been shown experimentally⁴⁴ that participation in collective action — specifically, voting in a national election — was more likely when the subjects were made to think of themselves as typical members of an activist group whose collective behaviour would decide the outcome. Each individual would then reason in the following manner: “I am a fairly typical member of my group. If I vote, others are likely to vote as well. Being like me, they will tend to act like me. Let me vote, therefore, to bring it about that others also vote and our party wins.” Magical thinking is at the root of everyday Kantianism.

⁴² BANFIELD (1958).

⁴³ QUATTRONE and TVERSKY (1986).

⁴⁴ QUATTRONE and TVERSKY (1986).

The norm of fairness cannot by itself induce cooperation, although it can act as a multiplier on cooperation once it has arisen in other people for other reasons. We may imagine a scenario in which the first cooperators are everyday Kantians. Later they are joined by utilitarians, who need Kantians to bring cooperation over the threshold where it begins to yield net benefits. Together, Kantians and utilitarians trigger off the norm of fairness, by providing a pattern of behaviour to which the conformists have to conform. I am not saying that this particular scenario is frequent, only that it is the kind of scenario needed for the norm of fairness to yield its benefits.

Hence, if we want to explain the norm of fairness by its collectively useful consequences, it must be as part of the explanation of a complex package of motivations. Perhaps it is more plausible to think of it as a combination of conformism and envy. Recall that a collective action problem is defined by the fact that cooperation has diffuse benefits and precise costs. In the standard argument, this provides individuals with a reason for defecting. But we can turn this argument on its head. It is precisely because contributions are easier to identify than their effects that they can become the object of a social norm to cooperate if and only if others do. Also, for many people the thought that others are getting more lightly off than themselves is hard to tolerate, even when the group would actually benefit from having some free riders. (The analogy to the maxim from Jewish law should be clear.)

Another reductionist strategy is to argue that social norms are expressions of individual rationality. They are optimal rules of thumb, designed to cope with the limitations of human nature. An outcome-oriented utilitarian might well decide to follow everyday Kantianism as a second-best solution. To apply first-best utilitarian principles one must estimate the number of cooperators and the technology of collective action. These estimates are costly, inherently uncertain, and subject to self-serving biases. In particular, it is difficult to treat the cost to oneself of cooperating merely as a parameter to be estimated. Knowing this, a rational utilitarian might well decide to abdicate from fine-tuning and to follow instead a rigid, inflexible rule of always cooperating. For Descartes, for instance, “la plus grande finesse est de ne vouloir point du tout user de finesse.”⁴⁵

Social norms against drinking, gambling, smoking, and the like could

⁴⁵ DESCARTES (1897–1910: vol. 4, p. 357).

be understood in a similar perspective.⁴⁶ It is not necessarily in the long-term self-interest of the individual to abstain wholly from these activities. At moderate levels, the damage they do to the body or to the purse can be offset by the pleasures they bring, just as the benefits to other people of my cooperation can be offset by the cost to me of providing them. Yet compromises are not always feasible. For physiological and psychological reasons, the choice may be between engaging in these activities at a high and destructive level and not engaging in them at all. A rational individual who is aware of this fact might well decide to adopt an inflexible rule, "never suffer a single exception". Although really a "private law", it appears as a social norm because it is shared by many individuals and because people try to inculcate it in others they care about.

This strategy suffers from two flaws. First, it clearly cannot account for all social norms. Norms of vengeance, for instance, are individually as well as collectively irrational. It is better for all if none follow them than if all do. It is also better for the individual not to follow them if others do. (By this I mean, of course, better if we disregard the shame and dishonour attached to their violation.) That it might be better to follow them if others do not is not a relevant consideration in this context, since there is no such thing as individual adherence to a social norm. More generally, the strategy does not work well for metanorms. The norm to avoid incest may be individually rational, but the same does not seem to be true of the norm to criticize incest.

Secondly, the strategy ignores the fact that people are *in the grip* of social norms, and that they work because and to the extent that people are in their grip. Violation of norms generates anxiety, guilt and shame, together with their physiological concomitants blushing, sweating and the like. Social norms are mediated by emotions, which are triggered by actual or anticipated sanctions by other people or by the call from one's own conscience. Although the rational character planner might want to create the appropriate emotional reaction patterns in himself, they are only marginally within his control.

This remark brings me to the third reductionist strategy, which tries to explain social norms in terms of genetic fitness. The emotions of guilt and shame that are needed to buttress social norms might be the product of

⁴⁶ For elaborations of the following argument, see AINSLIE (1982, 1984, 1986) and ELSTER (1985).

natural selection. Norms against incest, female promiscuity, myopic behaviour or cheating in cooperative relationships might conceivably be explained by their fitness-enhancing effect. This strategy could be used to support the second, individualist variety of reductionism, but it could also work against it. Robert Trivers's explanation of cooperative behaviour in terms of reciprocal altruism does not violate individual optimality, whereas William Hamilton's explanation of cooperation in terms of inclusive fitness and kin selection does require suboptimal behaviour of the individual.⁴⁷

I do not have the space or competence to go deeply into the sociobiological controversies. Let me nevertheless, building on Philip Kitcher's superb book *Vaulting Ambition*⁴⁸, offer some remarks on the inadequacies of genetic reductionism.

First, we should firmly resist the temptation to think that any pervasive and stable item of human behaviour must be fitness-enhancing. Shame is a universal emotion which probably has a genetic basis. It might conceivably owe its existence to its ability to stabilize fitness-enhancing behaviour. It might also, however, be a pleiotropic by-product of genes that have been favoured for other reasons. I cannot see how one could even begin to establish that shame, taken in isolation, has net beneficial effects. It may favour cooperation, but also conformism, conservatism and self-destructive codes of honour. On first principles it seems likely that the tendency to feel shame is part of some fitness-optimizing package, but there is no reason to believe that every single item in the package is optimal.

Next, we should resist the temptation to tell plausible-sounding stories. Story-telling has its place, both in the social and in the biological sciences, which is to refute *a priori* anti-reductionism. One writer may maintain that item X *cannot possibly* be explained by a theory of type T. Cooperation among firms cannot possibly be explained as a rational strategy for individual profit-maximizing. Avoidance of overgrazing among animals cannot possibly be explained as an optimal strategy for reproductive fitness. Both statements can be refuted by telling a plausible story, deducing cooperative behaviour from individual rationality or fitness. Plausibility does not amount to truth, however. The story, to be plausible, must assume specific values of parameters about which we are

⁴⁷ TRIVERS (1971), HAMILTON (1964).

⁴⁸ KITCHER (1985).

often ignorant. The assumptions are legitimate and useful in refuting *a priori* anti-reductionism, but they do not establish reductionism.

This point is a quite general one, that also applies to the first two reductionist strategies. Any economist worth his salt could tell a story that would show either the collective or the individual rationality of the norm against buying places in the queue or of the norm of vengeance. The Jalé norm of strict liability could be seen as a rational response to problems of information and incentive compatibility. The very ease with which a practiced social scientist can invent such stories should warn us against believing in their explanatory power.

Thirdly, we should resist the temptation to seek a genetic foundation for specific social norms. In some societies there are norms against corruption and against those who do not denounce corruption. In other societies there are no norms against corruption, but norms against those who denounce it. What these societies have in common is the presence of norms which guide behaviour. The tendency to be guided by social norms probably has a genetic explanation, but, to repeat my first warning, there is no reason to think that the tendency is optimal in any other sense than being part of an optimal package of tendencies to act and to react.

This paper has no punchline. I have no answer to the riddle of social norms: their origin, maintenance, change and variation. Social norms may have something to do with fundamental desires of human beings: the desire to be like others, the desire for others to be like oneself, and the desire to differ from others. They may stem from a need to simplify decisions and to avoid the tensions and stress of personal responsibility. Sometimes, we care more about making a clearcut decision than about making the right decision. Other conjectures and speculations could be offered, but I shall not do so here. Instead I want to reemphasize the main points of my paper. Social norms have causal efficacy. There is no reason to think that they are the outcome of some optimizing hidden hand.

References

- AINSLIE, G., 1982, *A behavioral economic approach to the defense mechanisms: Freud's energy theory revisited*, *Social Science Information* 21, pp. 735–779.
- AINSLIE, G., 1984, *Behavioral Economics II: Motivated involuntary behavior*, *Social Science Information* 23, pp. 247–274.
- AINSLIE, G., 1986, *Beyond microeconomics*, in J. Elster, ed., *The Multiple Self* (Cambridge University Press), pp. 133–175.

- AKERLOF, G., 1980, *A theory of social custom, of which unemployment may be one consequence*, Quarterly Journal of Economics, 94, pp. 749–776.
- ARROW, K., 1971, *Political and economic evaluation of social effects and externalities*, in M.D. Intriligator, ed., *Frontiers of Quantitative Economics* (North-Holland, Amsterdam), pp. 3–25.
- AXELROD, R., 1984, *The Evolution of Cooperation* (Basic Books, New York).
- AXELROD, R., 1986, *An evolutionary approach to norms*, American Political Science Review, 80, pp. 1095–1111.
- BANFIELD, E.G., 1958, *The Moral Basis of a Backward Society* (The Free Press, New York).
- BARRY, B., 1979, *Sociologists, Economists and Democracy*, 2nd edn. (Chicago University Press).
- BECKER, G., 1962, *Irrational behavior and economic theory*, reprinted as Ch. 8 of G. Becker, *The Economic Approach to Human Behavior* (University of Chicago Press, 1976).
- BOURDIEU, P., 1980, *La Distinction* (Editions de Minuit, Paris).
- CANCIAN, F., 1975, *What are Norms?* (Cambridge University Press).
- DESCARTES, R., 1897–1910, *Oeuvres Complètes*, ed. Adam and Tannery (Paris).
- DEUTSCH, M., *Distributive Justice* (Yale University Press, New Haven).
- EDGERTON, R.B., 1985, *Rules, Exceptions and Social Order* (University of California Press, Berkeley).
- ELSTER, J., 1983, *Sour Grapes* (Cambridge University Press).
- ELSTER, J., 1985, *Weakness of will and the free-rider problem*, Economics and Philosophy 1, pp. 231–265.
- ELSTER, J., 1986, *Editorial Introduction to Rational Choice* (Blackwell, Oxford).
- ELSTER, J., 1988, *Is there (or should there be) a right to work?*, in A. Guttman, ed., *Democracy and the Welfare State* (University Press, Princeton).
- ELSTER, J., (forthcoming), *When rationality fails*, in: K. Cook and M. Levi, eds., *Limits to Rationality* (to be published).
- FØLLESDAL, D., 1981, *Sartre on freedom*, in: P.A. Schilpp, ed., *The Philosophy of Jean-Paul Sartre* (Open Court, LaSalle, Ill), pp. 392–407.
- GAMBETTA, D., 1987, *Did They Jump or Were They Pushed?* (Cambridge University Press).
- GOULDNER, A., 1960, *The norm of reciprocity*, American Sociological Review 25, pp. 161–178.
- HAMILTON, W., 1964, *The genetic evolution of social behavior*, Journal of Theoretical Biology 12, pp. 1–52.
- KAHN, A., LAMM, H. and NELSON, R., 1977, *Preferences for an equal or equitable allocator*, Journal of Personality and Social Psychology 35, pp. 837–844.
- KITCHER, P., 1985, *Vaulting Ambition* (MIT Press, Cambridge, Mass).
- LAURIN, U., 1986, *På Heder och Samvete* (Norstedt, Stockholm).
- LEVY, R., 1973, *The Tahitians* (University of Chicago Press).
- MARSDEN, D., 1986, *The End of Economic Man?* (Wheatshaf Books, London).
- MESSICK, D.M. and SENTIS, K., 1983, *Fairness, preference, and fairness biases*, in D.M. Messick and K. Cook, eds., *Equity Theory* (Praeger, New York), pp. 61–94.
- NORTH, D., 1981, *Structure and Change in Economic History* (Norton, New York).
- PIDDOCKE, S., 1965, *The potlatch system of the Southern Kwakiutl*, Southwestern Journal of Anthropology 21, pp. 244–264.
- QUATTRONE, G. and TVERSKY, A., 1986, *Self-deception and the voter's illusion*, in: J. Elster, ed., *The Multiple Self* (Cambridge University Press), pp. 35–58.
- RICE, O., 1982, *The Hatfields and McCoys* (University of Kentucky Press, Lexington).
- ROSNER, F., 1986, *Modern Medicine and Jewish Ethics* (Yeshiva University Press, New York).

- SELTEN, R., 1978, *The equity principle in economic behavior*, in: H. Gottinger and W. Leinfellner, eds., *Decision Theory and Social Ethics* (Reidel, Dordrecht) pp. 289–301.
- SIMPSON, A.W.B., 1984, *Cannibalism and the Common Law* (University of Chicago Press).
- SOLOW, R., 1980, *Theories of unemployment*, *American Economic Review* 70, pp. 1–11.
- TAYLOR, M., 1987, *The Possibility of Cooperation* (University Press, Cambridge).
- THALER, R., 1980, *Towards a positive theory of consumer choice*, *Journal of Economic Behavior and Organization* 1, pp. 39–60.
- TRIVERS, R., 1971, *The evolution of reciprocal altruism*, *Quarterly Review of Biology* 46, pp. 35–57.
- TURNBULL, C., 1971, *The Mountain People* (Simon and Schuster, New York).
- ULLMANN-MARGALIT, E., 1977, *The Emergence of Norms* (Oxford University Press).
- VEYNE, P., 1976, *Le Pain et le Cirque* (Editions du Seuil, Paris).
- WALLER, W., 1937, *The rating and dating complex*, *American Sociological Review* 2, pp. 727–734.
- WALZER, M., 1983, *Spheres of Justice* (Basic Books, New York).
- WRONG, D., 1961, *The oversocialized conception of man in modern sociology*, *American Sociological Review* 26, pp. 183–193.

ON THE NATURE OF A SOCIAL ORDER

INGMAR PÖRN

Department of Philosophy, University of Helsinki, Helsinki, Finland

1. I can indicate my approach to the subject, in broad outline, by saying that I rely on three categories: control, influence, and normative regulation. Control, again in broad terms, is a matter of what an agent *does* in relation to another agent; influence a matter of what an agent *can* do in relation to another; and normative regulation a matter of what an agent *shall* or *may* do in relation to another.

More generally, control is not only a matter of what an agent does in relation to another but also a matter of what he does *not* do in relation to him. And similar remarks may be made about influence and normative regulation. So I shall speak of three classes of relations between two agents and a state of affairs, namely control positions, influence positions, and normative positions. I believe that these are of central importance for the understanding of a social order. I also believe that they can be systematically characterized and that the characterization may be given by means of tools available in logical theory and modal logic in particular.

2. I shall use some systems of modal logic. All of them are quite simple, either normal systems of type KT or type KD or else constructed from systems of such types.

The smallest normal system of modal logic—called K by CHELLAS (1980: ch. 4)—comprises all and only the theorems that may be obtained from

PL propositional logic

Df $\diamond p = \neg \Box \neg p$

RK
$$\frac{p_1 \& p_2 \& \cdots \& p_n \rightarrow p}{\Box p_1 \& \Box p_2 \& \cdots \& \Box p_n \rightarrow \Box p} \quad (n \geq 0)$$

System KT is got by adding

$$T \quad \Box p \rightarrow p$$

to K and system KD is obtained by adding

$$D \quad \Box p \rightarrow \Diamond p$$

to K.

KT-models are structures of the kind $M = (U, R, V)$ where U is a non-empty set (of possible worlds), R a binary reflexive relation in U , and V a valuation function which assigns a truth value to each atomic sentence for each world $u \in U$. Truth at a point in the model is defined in the usual way for propositional connectives and for $\Box p$ ('It is necessary that p ') by the condition

$$M \models \Box p[u] \text{ if and only if } M \models p[v] \text{ for every } v \in U \\ \text{such that } (u, v) \in R.$$

The corresponding condition for $\Diamond p$ ('It is possible that p ') is of course

$$M \models \Diamond p[u] \text{ if and only if } M \models p[v] \text{ for some } v \in U \\ \text{such that } (u, v) \in R.$$

KD-models are structures of the same kind as KT-models except that R is serial in U .

3. If we wish to use tools drawn from modal logic for the purpose of characterizing a minimal logic of action we may proceed in the following way. We define a system D_a of type KT in which $D_a p$ is read as "It is necessary for something which a does that p " and $C_a p =_{df} \neg D_a \neg p$ as "It is compatible with everything which a does that p ".

D_a -models are structures of the kind $M = (U, R_{D_a}, V)$ in which $(u, v) \in R_{D_a}$ if and only if everything which a brings about in u is the case in v .

We next define a system D'_a of type KD in which $D'_a p$ is read as "But for a 's action it would be the case that p " and $C'_a p =_{df} \neg D'_a \neg p$ as "But for a 's action it might be the case that p ". In D'_a -models, which are structures of the kind $M = (U, R_{D'_a}, V)$, $R_{D'_a}$ is such that $(u, v) \in R_{D'_a}$ if and only if not everything which a brings about in u is the case in v .

The two systems, D_a and D'_a , may be combined in several ways. A minimal logic of action results if we introduce certain definitions and one axiom. The definitions I have in mind are as follows:

$$\text{Df } E_a \quad E_a p = D_a p \ \& \ C'_a \neg p$$

$$\text{Df } F_a \quad F_a p = C_a p \ \& \ C'_a \neg p$$

$$\text{Df } \text{ActN}_a \quad \text{ActN}_a p = D_a p \ \& \ D'_a p$$

$$\text{Df } \text{ActM}_a \quad \text{ActM}_a p = \neg \text{ActN}_a \neg p$$

$E_a p$ should be read as “ a brings it about that p ” and $F_a p$ as “ a lets it be the case that p ”. $\text{ActN}_a p$ says that p is necessary for something which a does and without a 's action p would (still) be the case. It says, in other words, that p is such an action necessity for a which is independent of his action. Action necessity in this sense is, as we shall see, a very useful construction.

The axioms of systems D_a and D'_a are of course standard and may be obtained immediately from any axiomatization of KT and KD, respectively, by writing \square in KT as D_a and \square in KD as D'_a . The only additional axiom of the combined system—let us call it $D_a D'_a$ —is

$$\text{ActM}_a F_a p^* \quad (p^* \text{ is a propositional constant}) \quad (1)$$

or, alternatively, if we allow ourselves to use quantifiers to bind propositional variables, the axiom is

$$\exists p \text{ActM}_a F_a p . \quad (\text{Q1})$$

The axiom expresses the possibility of the weakest kind of action definable in $D_a D'_a$, namely action of type letting, and it is to be expected since we are talking about an individual capable of some agency.

Among the theorems provable in $D_a D'_a$ the following may be noted:

$$E_a p \rightarrow p \quad (2)$$

$$E_a (p \rightarrow q) \rightarrow (E_a p \rightarrow E_a q) \quad (3)$$

$$(E_a p \ \& \ E_a q) \rightarrow E_a (p \ \& \ q) \quad (4)$$

$$\neg E_a t \quad (t \text{ a tautology}) \quad (5)$$

$$E_a p \rightarrow F_a p \quad (6)$$

$$\text{ActN}_a p \rightarrow (E_a q \leftrightarrow E_a(p \& q)) \quad (7)$$

$$\text{ActN}_a p \leftrightarrow \text{ActN}_a(p \leftrightarrow t) \quad (8)$$

$$\text{ActN}_a(p \leftrightarrow q) \rightarrow (E_a p \leftrightarrow E_a q) \quad (9)$$

These theorems are self-explanatory. Formula (9) is essential in the analysis of inferences such as: since Mauno Koivisto is the President of Finland, a shakes hand with Mauno Koivisto if and only if a shakes hand with the President of Finland; or since b is a married man if and only if b is not a bachelor, a brings it about that b is a married man if and only if a brings it about that b is not a bachelor.

Among the rules of $D_a D'_a$

$$\frac{p \leftrightarrow q}{E_a p \leftrightarrow E_a q}$$

and its counterpart for F_a may be mentioned, and RK is of course valid for ActN_a .

$D_a D'_a$ -models are structures of the kind $M = (U, R_{D_a}, R_{D'_a}, V)$, where the components are as before and where, minimally, the relations are subject to the condition

$$R_{D_a} \cap R_{D'_a} = \emptyset.$$

Axiom (1), alternatively (Q1), is its syntactic counterpart.

By the logic of action we understand in the sequel a collection of systems of type $D_a D'_a$ for some non-empty set of agents and we use DD' to refer to this logic. (For a more detailed discussion of a logic of action which is essentially similar to DD' , in respect of effective action, see PÖRN (1977: ch. 1).) DD' is not a normal system of modal logic.

4. By means of sentences of the form $E_a p$ we may express *effective actions* with the individual a as an agent. Sentences $p(a)$ which exhibit the singular term a but which are not of the form $E_a p$ express an effective action with a as an agent only if they are equivalent to the corresponding sentences of the form $E_a p(a)$. (For elaboration and defence of this idea, see KANGER (1972: pp. 122–124) and PÖRN (1977: pp. 11–16).)

By means of the modalities

$$\begin{array}{cccc} E_a & E_a \neg & F_a & F_a \neg \\ \neg E_a & \neg E_a \neg & \neg F_a & \neg F_a \neg \end{array}$$

applied to a state of affairs we may express simple two-place *act positions*; they are two-place because they involve one agent and one state of affairs and they are simple in contradistinction to positions definable in terms of them and logical operations.

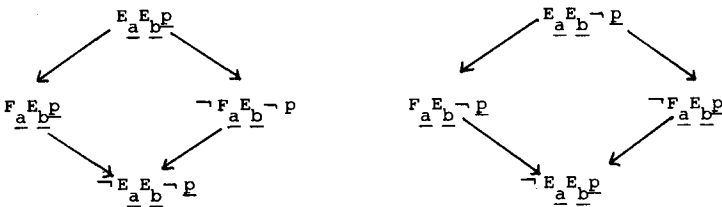
If the state of affairs of a two-place act position involves another individual *b*, as e.g. in “*a* walks on *b*’s land”, we have an act position *in relation to b*. Act positions in relation to another individual are of central importance for the understanding of social reality. Within the class of such positions we may delineate interesting subclasses. For example, using the modalities

$$\begin{array}{cccc}
 XY & XY\bar{} & X\bar{}Y & X\bar{}Y\bar{} \\
 \bar{}XY & \bar{}XY\bar{} & \bar{}X\bar{}Y & \bar{}X\bar{}Y\bar{}
 \end{array}$$

where

$$X = \begin{cases} E_a \\ F_a \end{cases} \quad \text{and} \quad Y = \begin{cases} E_b \\ F_b \end{cases}$$

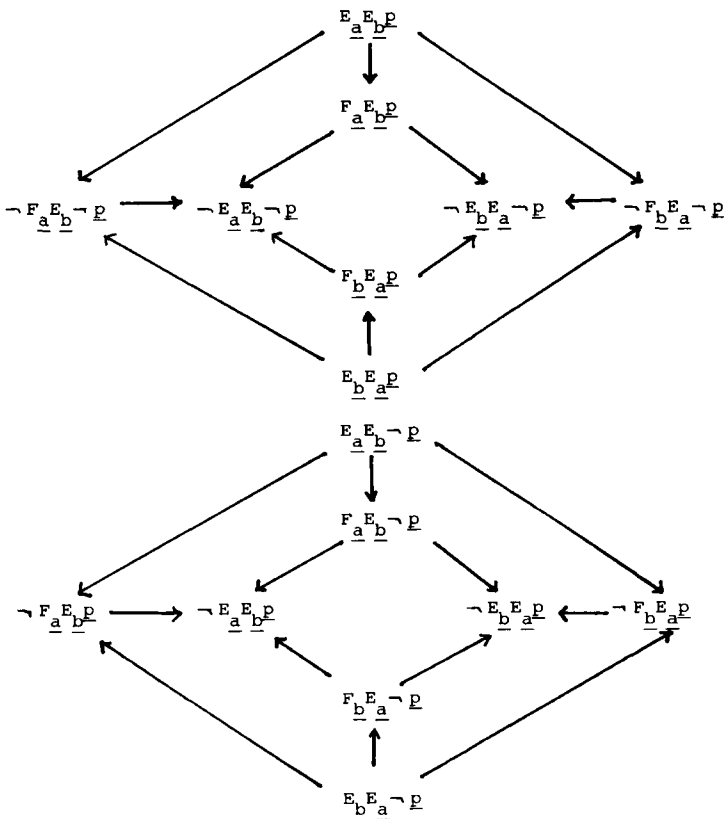
applied to a state of affairs *p* we obtain the subclass which comprise *a*’s act positions in relation to *b* with respect to *b*’s act positions with respect to *p*. In these cases, where *b* is referred to as an agent, I shall speak of *a*’s *control positions in relation to b* with respect to *p*. Within the class of control positions we may further define, e.g. the subclass comprising *a*’s control positions in relation to *b* with respect to *b*’s *effective* action with respect to *p*. In DD’ these positions are logically related as follows:



With the help of these diagrams we may readily ascertain the consistent constellations (conjunctions) of simple control positions with respect to the effective action of another individual. These—atomic two-place control positions—are summarized in the table below, in which + in a column indicates that a simple position appears and – that its negation appears as a component (conjunct) of a given constellation. Thus the rightmost column represents the atomic control position defined by the conjunction $\bar{}E_a E_b p$ & $\bar{}E_a E_b \bar{}p$ & $\bar{}F_a E_b p$ & $\bar{}F_a E_b \bar{}p$.

$E_a E_b p$	+	-	-	-	-	-
$E_a E_b \neg p$	-	+	-	-	-	-
$F_a E_b p$	+	-	+	+	-	-
$F_a E_b \neg p$	-	+	+	-	+	-

5. Having defined the class of (simple) two-place act positions, we may proceed to the characterization of three-place act positions. This is obtained, in the most general case, by combining, in accordance with DD', *a*'s act positions with respect to some state of affairs with *b*'s act positions with respect to the same state of affairs. Interesting subclasses may be defined. For example, for the analysis of aspects of interaction, the class of *intercontrol positions* is useful, i.e. the class of consistent conjunctions of *a*'s control positions in relation to *b* and *b*'s control positions in relation to *a* with respect to one and the same state of affairs. If, again, we restrict ourselves to control with respect to effective action, we are left with 16 positions which in DD' are logically related as follows.



There are 26 consistent conjunctions of the positions appearing in the diagrams—every other *prima facie* conjunction may be shown to be inconsistent by means of the theorems of DD' included in the diagrams. Moreover, the conjunctions are maximal. For illustration we consider

$$\neg E_a E_b p \ \& \ \neg E_a E_b \neg p \ \& \ F_a E_b p \ \& \ F_a E_b \neg p \ \& \\ \neg E_b E_a p \ \& \ \neg E_b E_a \neg p \ \& \ \neg F_b E_a p \ \& \ \neg F_b E_a \neg p .$$

This conjunction is maximal for the addition of a position from the class of 16 positions is either redundant or else it results in a contradiction. The conjunction may be shortened to

$$F_a E_b p \ \& \ F_a E_b \neg p \ \& \ \neg F_b E_a p \ \& \ \neg F_b E_a \neg p$$

since this conjunction logically implies the longer one. An instantiation might be:

a lets *b* bring it about that he (*b*) walks on his land & *a* lets *b* bring it about that he does not walk on his land & *b* does not let *a* bring it about that he (*b*) walks on his land & *b* does not let *a* bring it about that he does not walk on his land.

If a set of (relevant) states of affairs is given, we may try to determine, for each state of affairs in the set, the position, of the 26 possible intercontrol positions, which holds between *a* and *b* in respect of the state of affairs concerned. In this way we get the intercontrol profile of *a* and *b* with respect to the set of states of affairs. The intercontrol profile is one component of a more comprehensive story—the social narrative—of the large field of complex relations between *a* and *b*.

The method, here employed, of giving orderly accounts of relations between two parties was first used by KANGER (1963) in his explication of the concept of a right. The most impressive results of this line of development are to be found in KANGER and KANGER (1966). Their 26 rights-types are the obvious prototypes of the above intercontrol positions. The method has been elaborated and further developed by LINDAHL (1977) in his study of legal positions and in PÖRN (1970, 1977) the method is applied in the analysis and systematic description of control and influence.

6. Logic of action of the kind I have outlined makes no mention of the means that an agent uses to bring about a result. In case it be felt that this

is a defect, let me mention that system DD' can be extended by the addition of a *generation operation* \Rightarrow which at the basic level closely resembles the generation relation of GOLDMAN (1970). $p \Rightarrow q$ may be read “ p leads to q ” or “ q is a consequence of p ”. The extension is secured if we introduce the definition

$$E_a(p, q) =_{df} E_a p \ \& \ (p \Rightarrow q)$$

to capture the meaning of the construction “By bringing it about that p , a brings it about that q ”. SANDU (1986) has axiomatized the resulting dyadic logic of action and proved its soundness and completeness relative to the class of intended semantic models. It is obvious that such a logic of action supplies instruments for the description of intercontrol positions which are more refined than those available in DD' or, indeed, any monadic logic of action.

7. For the analysis of influence positions I suggest that we use a modal system G of type KT. In this $\Box p$ is written as

$$G_a p: \text{it is unavoidable for } a \text{ that } p$$

and $\Diamond p$ as

$$H_a p (=_{df} \neg G_a \neg p): \text{it is possible for } a \text{ that } p.$$

System G may be added to DD' to obtain DD'G. DD'G-models are structures of the kind

$$M = (U, R_{D_a}, R_{D'_a}, R_{G_a}, V)$$

for each agent a in the presupposed set of agents. In M all the components except R_{G_a} are as before and R_{G_a} is a reflexive relation in U such that $(u, v) \in R_{G_a}$ if and only if an action alternative open to a in u is realized in v . When R_{G_a} is understood in this way it is obvious that

$$R_{D_a} \subseteq R_{G_a}$$

and

$$R_{D'_a} \subseteq R_{G_a}.$$

The syntactic counterparts of these conditions are, respectively,

$$G_a p \rightarrow D_a p \tag{10}$$

and

$$G_a p \rightarrow D'_a p . \tag{11}$$

Some theorems of interest are:

$$G_a p \rightarrow \text{Act} N_a p \tag{12}$$

$$F_a p \rightarrow (H_a p \ \& \ H_a \neg p) \tag{13}$$

$$G_a p \rightarrow \neg E_a p \tag{14}$$

$$H_a \neg E_a p \tag{15}$$

$$E_a p \rightarrow H_a E_a p \tag{16}$$

How can influence be characterized within DD'G? If we replace E and F in left positions in the diagrams in section 5 by, respectively, G and F, the resulting implications are of course theorems of DD'G. Perhaps it may be said that the positions which these theorems link are positions of influence of a sort. However, they are not positions of influence of the kind which essentially involves the notion of a power to act in relation to another. $H_a E_b p$ only says that $E_a p$ is true in a world in which an action alternative open to a is realized. This might well be the case although a has no power in relation to b in the sense of having power to bring it about that $E_b p$. If, as I suggested in section 1, such a power is characteristic of a 's influence in relation to b then $H_a E_b p$ does not express it; it is too weak. An agent's power—with respect to effective action—must be articulated as $H_a E_a p$.

If we restrict ourselves to a 's effective action the modalities

$$\begin{array}{cccc} G_a E_a & G_a E_a \neg & G_a \neg E_a & G_a \neg E_a \neg \\ \neg G_a E_a & \neg G_a E_a \neg & \neg G_a \neg E_a & \neg G_a \neg E_a \neg \end{array}$$

are the *prima facie* candidates for simple power positions with respect to some state of affairs. However, in view of theorem (15) the first two must be excluded. Since their negations are logical truths they must also be

excluded—they do not serve to characterize a 's power in any substantial way. There remain, then, only four relevant modalities, namely

$$\neg G_a \neg E_a \quad \neg G_a \neg E_a \neg \quad G_a \neg E_a \quad G_a \neg E_a \neg$$

or, equivalently,

$$H_a E_a \quad H_a E_a \neg \quad \neg H_a E_a \quad \neg H_a E_a \neg$$

which, when applied to a state of affairs, yield four simple power positions. They combine, of course, to four atomic power positions, for one agent and one state of affairs. Similar results, within different frameworks, may be found in PÖRN (1977: §32) and KANGER (1977).

The above reasoning results in a powerful reduction of combinatorial possibilities in the case of power with respect to effective action. There does not seem to be a similar line of reasoning resulting in a corresponding reduction in the case of action of type letting. This field is therefore rich. And the richness is further increased if the two types of power are considered together. But I hasten to add that a considerable reduction may be possible in the latter field depending on how the notion of an action alternative open to an agent is articulated. This notion is presupposed in the characterization of R_{G_a} , but, unfortunately, it is rather vague; its explication is a desideratum if the study of power along the lines here suggested is to be further advanced.

An agent's power to act in relation to another may be called *social* power. Such a power is present if, for example, a can walk on b 's land. One reasonable interpretation of influence is obtained, it seems to me, if we make influence a subcategory of social power by defining influence as power with respect to another agent's effective action or action of type letting. There is an important difference between a 's power to walk on b 's land and a 's power to bring it about that b walks on c 's land, or a 's power to let b walk on c 's land, or a 's power to bring it about that b lets c walk on his land, etc.

Control and influence may be "combined". By this I mean that we may define control in relation to an agent with respect to his influence positions (with respect to some set of relevant states of affairs) and, conversely, his influence in relation to an agent with respect to his control positions. The studies of OPPENHEIM (1961, 1981) clearly show that work in these areas are important for the understanding of the foundations of social science. Other "combinations" must also be mapped. KANGER

(1977) shows that influence with respect to influence may have an interesting logic.

8. There is a close structural affinity between act positions and normative positions. For the latter we may rely, to begin with, on classical deontic logic, a system of type KD which I shall call O in this context. In system O , $\Box p$ is written as

$O p$: it is the case that p in every ideal world

and $\Diamond p$ as

$P p$ ($=_{df} \neg O \neg p$): it is the case that p in some ideal world.

O -models are structures of the kind $M = (U, R_O, V)$ where R_O is a serial relation in U such that $(u, v) \in R_O$ if and only if everything which shall be the case in u is the case in v . If $(u, v) \in R_O$, v may be said to be ideal relative to u ; hence the above readings of $O p$ and $P p$.

In JONES and PÖRN (1985, 1986) it is argued that we require, in a deontic logic less open to paradoxes than classical deontic logic, an operator by means of which we can describe what is the case in *sub-ideal* worlds, worlds in which something has gone wrong. This idea may be captured by the introduction of a system O' of type KD in which $\Box p$ is written as

$O' p$: it is the case that p in every sub-ideal world

and $\Diamond p$ as

$P' p$ ($=_{df} \neg O' \neg p$): it is the case that p in some sub-ideal world.

In O' -models $R_{O'}$ is a serial relation in U such that $(u, v) \in R_{O'}$ if and only if not everything which shall be the case in u is the case in v . If $(u, v) \in R_{O'}$, then v may be said to be sub-ideal relative to u ; v is indeed in that case a world in which something has gone wrong.

System OO' incorporates both systems, O and O' , and the following definitions, which are analogous to the definitions of four action operators in section 3:

Df Shall	Shall $p = Op \ \& \ P' \neg p$
Df May	May $p = Pp \ \& \ P' \neg p$
DeonN	DeonN $p = Op \ \& \ O'p$
DeonM	DeonM $p = \neg DeonN \neg p$

OO' also contains

$$\text{DeonM May } q^* \quad (q^* \text{ a propositional constant}) \quad (17)$$

and

$$\text{DeonN}p \rightarrow p \quad (18)$$

as axioms. Formula (17) is the deontic counterpart of axiom (1) of DD'G; in the quantifier version of OO' its place is taken by

$$\exists p \text{ DeonM May } p . \quad (\text{Q18})$$

The corresponding model conditions are, respectively,

$$\text{C1} \quad R_o \cap R_{o'} = \emptyset$$

and

$$\text{C2} \quad \Delta_U \subseteq R_o \cup R_{o'}$$

where Δ_U is the diagonal in U , i.e. $\Delta_U = \{(u, u) : u \in U\}$. As regards the effect of C1 there is here a departure from JONES and PÖRN (1985: pp. 278–279) and JONES and PÖRN (1986: p. 91).

Like DD', OO' is not a normal system of modal logic although the ingredients are normal. The DeonN-fragment of OO' is a system of type KT and as such normal. The OO'-counterparts of theorems (3)–(9) are valid in OO'. The analogue of (2) is of course

$$\text{Shall } p \rightarrow \text{May } p . \quad (19)$$

If E_a and E_b in left positions in the diagrams in section 5 are replaced by Shall and F_a and F_b in left positions by May, the resulting implications are theorems of OO'. If we restrict ourselves to normative regulation of

effective action, there are, therefore, 26 maximal consistent conjunctions of simple normative positions, for two agents and one state of affairs. For further details and systematic aspects of these matters, the reader is referred to KANGER and KANGER (1966) and LINDAHL (1977). As I have already indicated, their theory of normative positions was the blueprint for my treatment of intercontrol positions. It should also be noted in this context that their deontic logic is normal and therefore differs from the one suggested here.

If system OO' is added to DD'G, DD'GOO' results. This is a rather complex system whose complete characterization cannot be attempted here. I must leave open the question of how, for example, the following theses should be treated:

$$\text{Shall } E_a p \rightarrow H_a E_a p$$

$$\text{Act}N_a(p \leftrightarrow q) \rightarrow (\text{Shall } E_a p \rightarrow \text{Shall } E_a q)$$

$$G_a \text{Act}N_a(p \leftrightarrow q) \rightarrow (\text{Shall } E_a p \rightarrow \text{Shall } E_a q)$$

$$G_a F_a p \rightarrow \neg \text{May } E_a \neg p .$$

When modalities which appear to be simple when considered in isolation are allowed to appear in nested constructions perplexities soon arise.

9. If we wish to create a logically oriented framework for social narratives, something like DD'GOO' seems to be a minimum requirement. Social narratives describe relations between agents. These relations may be extremely complex, with respect to both types and features within a type. In practice a selection of types and features is therefore made.

By a *social order* narrative I shall here understand a social narrative which specifies the positions of control, influence and normative regulation for some agents. The unity of a social order is furnished by its normative elements; these constitute the kernel of the social order. More precisely, the kernel of a social order comprises a set of *position rules*. Control and influence enter the social order as interagent relations in which position rules are *realized* or else fail to be realized.

Normative positions appear in position rules. Such a rule (in standard form) assigns an atomic normative position to a condition. If $R(x, y, p)$ is an atomic normative position of x and y with respect to p and $Q(x, y)$ is the rule condition then the sentence

$$\text{Deon}N\forall x\forall y (Q(x, y) \rightarrow R(x, y, p))$$

may be used to express the rule. The application of the rule to the case $Q(a, b)$ yields the position $R(a, b, p)$ since the rule and the case logically imply $R(a, b, p)$.

If the case $Q(a, b)$ obtains, we say, following KANGER (1985), that the rule is realized for a in relation to b if the following conditions are satisfied:

- (i) $E_b p$ ($E_b \neg p$) if the application of the rule to the case yields Shall $E_b p$ (Shall $E_b \neg p$),
- (ii) $\neg E_b p$ ($\neg E_b \neg p$) if the application of the rule to the case yields \neg May $E_b p$ (\neg May $E_b \neg p$),
- (iii) $H_a E_a p$ ($H_a H_a \neg p$) if the application of the rule to the case yields May $E_a p$ & \neg May $E_b \neg p$ & \neg Shall $E_b p$ (May $E_a \neg p$ & \neg May $E_b p$ & \neg Shall $E_b \neg p$).

For example, if the application of the rule to the case implies each of

$$\text{May } E_a p \quad \text{May } E_a \neg p \quad \neg \text{May } E_b p \quad \neg \text{May } E_b \neg p$$

then the rule is realized for a in relation to b if

$$H_a E_a p \quad H_a E_a \neg p \quad \neg E_b p \quad \neg E_b \neg p .$$

It should be noted, in this case, that the rule may be realized although the actual intercontrol profile includes $\neg F_b E_a p$ or $\neg F_b E_a \neg p$.

In KANGER (1984) realization of the position rules embedded in the U.N. Declaration of Human Rights is studied in detail. In LINDAHL (1980) realization of position rules is considered as a problem of distribute justice.

The concept of realization is important for the understanding of other aspects of position rules. For this we may take protection as an example. In normally complex normative orders secondary position rules are superimposed on primary rules. HART (1961: ch. V) subdivides the secondary rules of a *legal* order into rules of recognition, rules of change, and rules of adjudication. For the protection of primary rules it is necessary that secondary rules be added to them. This is not enough, however. The important consideration, for protection, is of course the question whether or not the secondary rules concerned are realized. Thus, if we take the possible complexity of the normative kernel of a social order into account, the realization and protection of its primary

rules will be seen to give the social order a correspondingly richer texture of influence and control.

References

- CHELLAS, B.F., 1980, *Modal Logic: An Introduction* (Cambridge University Press, Cambridge).
- GOLDMAN, A.I., 1970, *A Theory of Human Action* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey).
- HART, H.L.A., 1961, *The Concept of Law* (Oxford University Press, Oxford).
- JONES, A.J.I. and PÖRN, I., 1985, *Ideality, Sub-Ideality and Deontic Logic*, *Synthese* 65, pp. 275–290.
- JONES, A.J.I. and PÖRN, I., 1986, 'Ought' and 'Must', *Synthese* 66, pp. 89–93.
- KANGER, H., 1984, *Human Rights in the U.N. Declaration* (Acta Universitatis Upsaliensis, Skrifter utgivna av Statsvetenskapliga föreningen i Uppsala nr 97, Uppsala).
- KANGER, S., 1963, *Rättighetsbegreppet*, in: Sju filosofiska studier tillägnade Anders Wedberg den 30 mars 1963 (Philosophical Studies published by the Department of Philosophy, University of Stockholm, No. 9, Stockholm).
- KANGER, S., 1972, *Law and Logic*, *Theoria* 38, pp. 105–132.
- KANGER, S., 1977, *Några synpunkter på begreppet inflytande*, in: *Filosofiska smulor tillägnade Konrad Marc-Wogau* (Filosofiska studier utgivna av Filosofiska institutionen vid Uppsala Universitet, Uppsala).
- KANGER, S., 1985, *On Realization of Human Rights*, in: G. Holmström and A.J.I. Jones, ed., *Action, Logic and Social Theory* (Acta Philosophica Fennica 38).
- KANGER, S. and KANGER, H., 1966, *Rights and parliamentarism*, *Theoria* 32, pp. 85–115.
- LINDAHL, L., 1977, *Position and Change. A Study in Law and Logic* (D. Reidel Publishing Company, Dordrecht).
- LINDAHL, L., 1980, *Fördelningsrättvisa och mänskliga rättigheter*, in: H. Kanger, L. Lindahl and M. Sjöberg, eds., *Fördelning rättvisa och mänskliga rättigheter* (Reports of the Research Project Models For Justice No 1980:1, Department of Philosophy, University of Uppsala, Uppsala).
- OPPENHEIM, F.E., 1961, *Dimensions of Freedom* (St Martin's Press, New York).
- OPPENHEIM, F.E., 1981, *Political Concepts* (Basil Blackwell, Oxford).
- PÖRN, I., 1970, *The Logic of Power* (Basil Blackwell, Oxford).
- PÖRN, I., 1977, *Action Theory and Social Science* (D. Reidel Publishing Company, Dordrecht).
- SANDU, G., 1986, *Formal Logic of Action*, Licentiate Thesis, University of Helsinki, Helsinki.

This Page Intentionally Left Blank

12

Foundations of Linguistics

This Page Intentionally Left Blank

INFORMATIONAL INDEPENDENCE AS A SEMANTICAL PHENOMENON

JAAKKO HINTIKKA and GABRIEL SANDU

*Department of Philosophy, Florida State University, Tallahassee, FL 32306, USA and
Department of Philosophy, University of Helsinki, Helsinki, Finland*

1. The concept of informational independence

Many linguists and philosophers of language may have heard of informational independence, but most, not to say virtually all, of them consider it as a marginal feature of the semantics of natural languages. Yet in reality it is a widespread phenomenon in languages like English. In this paper, we shall develop an explicit unified formal treatment of all the different varieties of informational independence in linguistic semantics. This treatment amounts to a new type of logic, which is thereby opened for investigation. We shall also call attention to several actual linguistic phenomena which instantiate informational independence and provide evidence of its ubiquity. Last but not least, we shall show that the phenomenon of informational independence prompts several highly interesting methodological problems and suggestions.

The concept of informational independence (II) belongs to game theory and it is applicable to logical and linguistic semantics in so far as that semantics can be dealt with by means of game-theoretical conceptualizations.¹ For this reason, any success that this concept might have as an explanatory tool provides further evidence for game-theoretical semantics (GTS).²

The concept of II contains essentially just what one would expect upon hearing the term. In games like chess, each player has access to the entire

¹ Game theory was created by VON NEUMANN and MORGENSTERN (1944). For recent expositions, see, e.g. OWEN (1982) or JONES (1980).

² GTS is an approach to logical and linguistic semantics developed by Jaakko Hintikka and his associates. For it, see SAARINEN (1979), HINTIKKA (1983), HINTIKKA and KULAS (1985), and HINTIKKA (1987a).

earlier history of the game, but in many others a player's knowledge of what has happened earlier is incomplete. In this case we are dealing with a game with imperfect information. A move made in ignorance of another one is said to be informationally independent of the latter. For instance, in many card games one does not know which cards one's opponent has picked up earlier.

It is not difficult to see how informational independence (and dependence) can be handled in general. In the mathematical theory of games, a game is represented (in its extensional form) by a labelled tree whose elements are all the possible situations in which one of the players makes a move. Each such situation comes with an *information set* which shows which other moves the player in question is aware or unaware of in making the move, i.e. which other moves the present one is dependent on or independent of.

2. Informational independence and the concept of scope

In GTS, certain games, called semantical games, play a crucial role in the analysis of the semantics of natural languages. Hence, the concept of II can be used without further ado in GTS. More than that, the very possibility of defining II brings out to the open several important restrictive presuppositions which are all too generally made in linguistics.

Even though the concept of II may be a novelty to many linguists, its twin, the concept of informational dependence, is an important stock in trade of all logically oriented linguists. It is one of the things that are dealt with by means of the ubiquitous concept of *scope*. For what is it that the concept of scope does in semantics? For instance, what does it really mean that in (2.1) "someone" has a wider scope than "everybody" but that in (2.2) this relation is reversed?

Someone loves everybody . (2.1)

Everybody is loved by someone . (2.2)

You do not need to be steeped in the technicalities of GTS to appreciate the role of informational dependence in (2.1) and (2.2). (2.1) is true if you are able to find a lover such that whoever else is chosen from the universe of discourse turns out to be one of your chosen lover's *inamorata*. In contrast, in (2.2) you only have to be able to find a lover

for whoever an imaginary opponent might choose from the relevant domain of individuals. In other words, in (2.1) your choice of a lover is independent of the choice of any one of his or her loved ones, whereas in (2.2) it may depend on the latter choice.

In brief, what this example shows is that a part of what *being within the scope of* means is *being informationally dependent on*. This is a part of the cash value of the notion of scope. But as soon as we see this, we can see that the received notion of scope is a hopeless mess.³ For it presupposes a nested ordering of the scopes of the different logically active ingredients of a sentence. There is no earthly reason for this assumption, which is tantamount to assuming that “the game of language” is a game with perfect information. On the contrary, there is plenty of evidence that this is not the case universally.⁴

Thus we realize two things: (i) one of the functions of the received notion of scope is to indicate the information sets of different moves in GTS; and (ii) the received notion of scope is a bad way of doing so, for it excludes arbitrarily certain empirically possible phenomena.

Once we realize all this, we can also see that the traditional notion of scope has other functions and other prejudices built into it. For in the usual Frege–Russell notation the scope of various logical operators does not only indicate their respective logical priorities. For this purpose, no brackets would be needed. All we would have to do is to number (index) the different operators (and other relevant ingredients of a sentence). But in the conventional scope notation, something else is done. To each quantifier a chunk of a sentence (or a text) is associated, usually by means of a pair of brackets, in which certain variables are supposed to be bound to a quantifier. Whatever there is to be said of this function of the notion of scope, it is different from, and independent of, the task of indicating logical priorities. This can be illustrated by means of the following examples, one of which is ill-formed in the usual notation but nonetheless makes perfect sense.⁵

$$P\{(\forall x)(A(x)) \supset B(x)\} \quad (2.3)$$

³ See here HINTIKKA (1987b).

⁴ Such evidence will be presented in the course of this paper. The first ones to show the presence of II in natural languages in forms other than partially ordered quantifiers were CARLSON and TER MEULEN (1979).

⁵ The example (2.3) goes back at least to David Kaplan. We do not know whether he ever published it, however.

$$P\{(\forall x)(A(x) \supset B(x))\} \quad (2.4)$$

$$(\forall x)(P\{A(x)\} \supset B(x)) \quad (2.5)$$

Here “ P ” is the possibility-operator, i.e. a kind of quasi-quantifier which says “in at least one alternative world it is the case that”. Hence (2.4) says that in some alternative world all A 's are B 's, while (2.5) says that all individuals that are possibly A 's are in fact B 's. (2.3) is ill-formed, and yet it has a clear import. It says that there is at least one alternative world such that, whatever is an A , there is in fact (in the actual world) a B . This is different from what (2.4) and (2.5) say, and it cannot be expressed by any expression of the conventional modal logics.

For our present purposes, the relevance of (2.3)–(2.5) lies in showing that the two component functions of the notion of scope are independent of each other. In (2.3) and (2.4) the logical priorities of “ P ” and “ $(\forall x)$ ” are the same, but the segments of the formula that constitute their respective “scopes” are different. In (2.3) the *Wirkungsbereich* of both “ P ” and “ $(\forall x)$ ” is the same as in (2.5), but their logical order is different.

It is not even clear whether the second function of the notion of scope (viz. that of indicating the syntactical limits of binding) makes any sense when applied to natural languages. This question is discussed in HINTIKKA (1987b).⁶

3. Partially ordered quantifiers and their implications

Among logicians, the phenomenon of II is best known in the form of partially ordered (p.o.) quantifiers, e.g. branching quantifiers.⁷ They have been studied in some depth by logicians. We shall not try to summarize the literature here. Suffice it to call your attention to some of the most striking results. One of them is the reduction of the decision problem for the entire second order logic (with standard interpretation) to the decision problem for branching quantifiers formulas.⁸ This result raises various interesting questions. Jaakko Hintikka has argued that all different

⁶ Partially ordered quantifiers were introduced in HENKIN (1961). For a partial bibliography on them, see HINTIKKA (1983), pp. 300–303.

⁷ See HINTIKKA (1974), pp. 170–171.

⁸ See HINTIKKA (1974), pp. 168–170; (1970), §9.

types of branching quantifier prefixes are present in English as the semantical forms of English quantifiers sentences.⁸ If so, the logical strength of English semantics is incredibly greater than that of first-order logic, which turns out to be a singularly restrictive paradigm in linguistic semantics as a framework of semantical representation.

The really interesting problem here is nevertheless not the shortcomings of first-order logic, but how natural language manages to reach the extra power that lifts its semantics beyond the reach of first-order logic. This question is made pertinent by the fact that natural languages do not employ anything like the paraphernalia of higher-order logic with its quantification over higher order entities, e.g. over properties and relations, over properties of relations and relations of properties, etc.⁹ Much of the interest of the phenomenon of II lies in the very fact it represents one of the ways in which natural languages transcend the power of first-order logic without employing explicit higher-order quantifiers. For the most striking feature of branching quantifiers is that they increase the logical power of first-order logic without any increase in its ontology. This seems to suggest that the contrast between first-order logic and second-order logic is in a certain sense artificial.¹⁰

4. The varieties of II and their logic

Partially ordered quantifiers have been studied in some depth outside the literature on GTS, albeit by means of game-theoretical concepts. One service GTS performs here is to show that branching quantifiers are but the tip of the independence iceberg. Indeed, it is patent that a given application of any game rule whatsoever can in principle be informationally independent of that of any other. Moreover, in the course of the development of a systematic theory of GTS, it is seen that game rules must in fact be associated with a wide variety of linguistic expressions besides quantifiers and propositional connectives, e.g. with modal words, tenses, genitives, intensional verbs (e.g., verbs for propositional attitudes), pronouns, certain prepositional constructions, and even proper names.¹¹ It is thus possible within GTS to extend the phenomenon of II

⁹ One way in which natural languages lift themselves above the first-order level is the use of plurals and plural quantifiers. They will not be discussed in this paper however.

¹⁰ The same conclusion has been argued for in PUTNAM (1971).

¹¹ See here especially HINTIKKA and KULAS (1985), pp. 22–27, 88–89, 94–98, 170–178, 187–188, etc.

beyond quantifiers, in order to cover linguistic expressions of all kinds. It is in fact GTS that suggest the prediction that the phenomenon of II should be found in operation among all these kinds of expressions. One of the purposes of this paper is to point out that this prediction is amply fulfilled by the evidence.

What we are going to do is to follow the practice of game theorists and make it possible to associate to each linguistic expression of the kind that prompts a move in a semantical game an indication of the information set of the correlated move. Since normal informational dependence (dependence, that is, on operators within the indicated scope of which an expression occurs) is the null hypothesis here, one handy way is to allow merely an indication of which expression a quantifier or other logical operator is exceptionally independent of. We propose to express this by writing " X/YZ " which means that the moves prompted by the expression X is informationally independent of the expressions Y and Z . Since different occurrences of many expressions, e.g. of propositional connectives, are not distinguished from each other typographically, we may have to attach subscripts to them in order to make independence relations unambiguous. (Cf. (5.12) below.)

When we do this in a first-order logic, we obtain a new logic which is even notationally much more flexible than the received first-order logic, including its modalized and intensionalized extensions. One of the messages of our paper is to recommend this new logic to linguists as a much better framework of semantical representation than the usual first-order logic or quantified intensional logic.

It is to be noted that the slash notation does not provide only a syntax for a formal language into which expressions of natural language exhibiting II can be translated. The explanation just given provides a fully adequate semantical interpretation for all the expressions in the new notation within the framework of GTS.

In the new logic, we can have formulas like the following:

$$(\forall x)(A(x)(\vee / \forall x)B(x)) \quad (4.1)$$

$$K_{\text{John}}B((a/K_{\text{John}})) \quad (4.2)$$

$$(\forall x)(\exists y/\forall x)R(x, y) \quad (4.3)$$

$$\sim(B_{\text{John}}/\sim)S \quad (4.4)$$

$$P(\forall x)(A(x) \supset (B/P)(x)) \quad (4.5)$$

With a little bit of creative imagination, you can thus formulate a tremendous variety of expressions which at first will look wild but which can all easily be seen to have a perfectly sensible semantical interpretation.

This new logic remains to be studied. Indeed, we have here an incredibly rich and powerful logic which nevertheless largely is virgin territory.

Admittedly, the new logic is so strong that it cannot be axiomatized in its entirety.¹² This does not mean, however, that linguistically important fragments of the new logic cannot be axiomatized.

Many of the formulas of the new logic are reducible to the conventional ones. For instance, the conventional equivalents of (4.1)–(4.4) are, respectively,

$$(\forall x)A(x) \vee (\forall x)B(x) \quad (4.6)$$

$$(Ex)(x = a \ \& \ K_{\text{John}}B(x)) \quad (4.7)$$

$$(Ey)(\forall x)R(x, y) \quad (4.8)$$

$$B_{\text{John}} \sim S \quad (4.9)$$

(Qualification: (4.2) and (4.7) are unproblematically equivalent only if it is assumed that *a* in fact exists.) There are, however, formulas in the new notation which cannot be reduced to a conventional first-order formula (cf. §5, below). A case in point is (4.5), which is equivalent with (2.3).

5. Neg-raising as an independence phenomenon

The ubiquity of the phenomenon of II is best demonstrated by means of case studies. The first one we shall offer here concerns the phenomenon of neg-raising. It has been discussed frequently in linguistics. It is exemplified by the following English sentence:

$$\text{Thomas does not believe that John is at home.} \quad (5.1)$$

¹² This follows from the fact that the decision problem for the logic of branching quantifiers is of the same order of difficulty as the decision problem for the entire second-order logic with standard interpretation. This result is proved in HINTIKKA (1974).

The usual colloquial force of (5.1) is not that of a contradictory of

Thomas believes that John is at home. (5.2)

Rather, (5.1) is normally taken to express the same as

Thomas believes that John is not at home. (5.3)

Usually the semantical mechanism which is in operation in neg-raising is treated in a line or two, and all attention is concentrated on the question as to when (under what conditions) neg-raising takes place. This investigative strategy is precisely the wrong way around. What is theoretically interesting and what admits of a sharp answer is not the “when” question but the “how” one. Only by answering the latter can we hope to have a handle on the former.

As the very name “neg-raising” shows, what is assumed to happen in this phenomenon is a change in the relative logical priority of “ B_{Thomas} ” and “ \sim ”. This assumption is wrong, as we have argued in a separate paper.¹³ Our thesis is that the *prima facie* logical form of (5.1), which is

$\sim B_{\text{Thomas}}(\text{John is at home})$, (5.4)

is in actual usage changed not to

$B_{\text{Thomas}} \sim (\text{John is at home})$, (5.5)

which is the logical form of (5.3), but to

$\sim(B_{\text{Thomas}}/\sim)(\text{John is at home})$ (5.6)

which can also be written as follows:

$B_{\text{Thomas}} \left. \begin{array}{l} \diagdown \\ \diagup \end{array} \right\} (\text{John is at home})$ (5.7)

If one reflects on the meaning of (5.7) for a moment, one will see that it is logically equivalent with (5.5). Making “ B_{Thomas} ” and “ \sim ” informa-

¹³ See SANDU and HINTIKKA (forthcoming).

tionally independent has thus the same effect in (5.1) as the change of the logical priorities among the two.

At first sight, there might not seem to be much additional explanatory force to be gained by saying that a negation and a belief operator are informationally independent in examples like (5.1), instead of saying that the two reverse their normal logical order. In fact, we obtain from our treatment a wealth of verifiable predictions which specify when the *prima facie* neg-raising will (or will not) take place and which differ from the results of merely taking the so-called neg-raising to be a permutation of “ \sim ” and “ B ”.

Consider, for instance, the following example:

Nobody does not believe that Homer existed. (5.8)

More colloquially expressed, (5.8) says more or less the same as

Nobody doubts that Homer existed .

The apparent logical form of (5.8) is

$$\sim(Ex) \sim B_x(Ey)(\text{Homer} = y) . \quad (5.9)$$

On the conventional treatments of neg-raising, there is no reason why it should not take place in (5.9). This would result in attributing the following logical form to (5.8):

$$\sim(Ex) B_x \sim (Ey)(\text{Homer} = y) . \quad (5.10)$$

What this says is

Nobody believes that Homer did not exist. (5.11)

Now obviously this is different from what (5.8) is normally taken to say, i.e. it is not an acceptable reading of (5.8). Hence the conventional account offers no explanation why (5.8) has the force it in fact does.

The natural way of extending the independence assumption to this case is to assume that the belief-operator is independent of both of the earlier occurrences of the negation-symbol, i.e. to assume that the force of (5.8)

is neither (5.9) nor (5.10) but

$$\sim_1 (Ex) \sim_2 (B_x / \sim_1 \sim_2)(Ey)(\text{Homer} = y) \quad (5.12)$$

where we have been forced to subscript the negation-signs in order to indicate which one the belief-operator is independent of. Now what is the force of (5.12)? In playing a game with (5.12) the choice of individual to be a value of x is made by nature, with the roles of the two players being exchanged in the sequel. But since B_x is independent of \sim_1 , this exchange of roles takes place only after the move connected with B_x . The choice of the world connected with B_x is likewise made by nature. Accordingly, the force of (5.12) is obviously that of

$$(\forall x)B_x \sim_1 \sim_2 (Ey)(\text{Homer} = y) \quad (5.13)$$

which is equivalent to (5.9).

Thus the independence treatment predicts correctly the absence of the *prima facie* neg-raising reading (5.10). This observation can be generalized in that the assumption of II offers a wide range of explanations as to when the *prima facie* permutation (like the step from (5.9) to (5.10)) is or is not possible.

6. Questions with an outside quantifier

Another instance of the phenomenon of informational independence in epistemic logic is provided by questions with an outside universal quantifier.¹⁴ They are illustrated by the following example:

$$\text{Whom does everyone admire?} \quad (6.1)$$

understood in the sense which can perhaps also be captured by

$$\text{Whom does one (a person) admire?} \quad (6.2)$$

¹⁴ They have been dealt with also (and more extensively) in HINTIKKA (1982a, 1982b).

The desideratum of this question is¹⁵

I know whom everyone admires . (6.3)

This desideratum has apparently only two main readings:

$(\exists y)K_1(\forall x)(x \text{ admires } y)$. (6.4)

$(\forall x)(\exists y)K_1(x \text{ admires } y)$. (6.5)

For “ $(\exists y)$ ” has to precede “ K ” in order for (6.1) to be a wh-question. Hence the only question is where the universal quantifier “ $(\forall x)$ ” goes.

The reading (6.4) does not interest us here. Its presupposition is so strong as to make this reading of (6.1) relatively rare in actual discourse. Hence (6.5) is apparently the only possible remaining reading here.

Let a reply to (6.2) on the reading (6.5) of its desideratum be “his or her eldest brother”. This brings about the truth of the following (in the mouth of the questioner):

$K_1(\forall x)(x \text{ admires } f(x))$ (6.6)

where $f(x) = \text{the eldest brother of } x$. But (6.6) is a conclusive answer to (6.1) only if the following are true (in the mouth of the questioner):

$(\forall x)(\exists z)(x = z \ \& \ (\exists t)K_1(z = t))$ (6.7)

$(\forall x)(\forall y)(y = f(x) \supset (\exists z)(z = y \ \& \ (\exists t)K_1(z = t)))$. (6.8)

This condition or conclusive answers means that I literally have to know who everybody and his (or her eldest) brother are. This is typically an unreasonable demand on the answers to (6.1). This shows that (6.5) is not always a viable reading of the desideratum (6.3) of (6.1), and certainly not its only viable reading besides (6.4).

Thus we seem to have reached an impasse. The reading (6.5) seems to be the only possible logical representation of (6.3) (barring (6.4), which

¹⁵ For the logical theory of questions and their answers, including the concepts in this theory (e.g., the concepts of desideratum and of presupposition), see ΗΙΝΤΙΚΚΑ (1975). Intuitively speaking, the desideratum of a normal direct question specifies the epistemic state of affairs which the questioner is trying to bring about.

we are not interested in). Yet this reading was seen not to capture what we actually mean by (6.3).

Here a recognition of the possibility of II allows us to escape the dilemma. For fairly obviously the right reading of (6.3) is

$$\left. \begin{array}{c} (\forall x) \rightarrow (Ey) \\ \swarrow \quad \searrow \\ K_1 \end{array} \right\} (x \text{ admires } y) \tag{6.9}$$

(Cf. HINTIKKA (1982).) In our new notation, (6.9) can be expressed as

$$K_1(\forall x)(Ey/K_1)(x \text{ admires } y) . \tag{6.10}$$

Admittedly, (6.9)–(6.10) can also be expressed by the following second-order formula

$$(Ef)K_1(\forall x)(x \text{ admires } f(x)) \tag{6.11}$$

This expression is not, however, reducible to any linear first-order equivalent in the traditional notation. For this reason the right reading (6.9)–(6.10) of (6.3) cannot be captured by the traditional notation of (first-order) epistemic logic.

The correctness of the reading (6.9)–(6.11) of (6.3) is shown, among other things, by the fact that it gives rise to the right conditions of conclusive answerhood. For instance, a functional reply “ $g(x)$ ” is a conclusive answer to (6.1) or this reading of its desideratum iff

$$(Ef)K_1(\forall x)(g(x) = f(x)) , \tag{6.12}$$

i.e. if the questioner knows which function $g(x)$ is. No knowledge of the argument values or function values is needed. This is in fact the right conclusiveness condition; it amounts to knowing what the function g is.

7. The *de dicto* vs. *de re* distinction and informational independence

Perhaps the most intriguing application of the idea of II is to the distinction between what are known as the *de dicto* and *de re* readings of

certain natural language sentences.¹⁶ For instance, suppose I say that Elmo believes something about the junior Senator from Florida (in short, about j), say.

$$\text{Elmo believes that } S[j] \quad (7.1)$$

Now two different things can be meant by (7.1). Either Elmo is said to believe something about whoever j might be, or else he is said to have a certain belief about a certain person who, possibly unbeknownst to our friend Elmo, is in fact the Junior Senator from the great state of Florida. In the present case, at the present moment, this means that Elmo believes something about Bob Graham the gentleman, without necessarily knowing or believing that he is in fact j .

The major puzzle about this distinction for the best philosophers of language was for a long time: Why should there be any problem whatever about the distinction? For any halfway reasonable model-theoretical treatment of the problem immediately shows what the distinction amounts to. On the *de dicto* reading, I am speaking of the (possibly different) individuals who in their respective scenarios (possible worlds, situations, or what not) play the role of the Junior Senator from Florida. In (7.1), these are the different scenarios compatible with everything Elmo believes. On the *de re* reading, we are taking the individual who in fact is the Junior Senator and following him throughout the same scenarios. What can be clearer than this explication of the contrast? It can in fact be expressed (or so it seems) in our usual notation of doxastic logic as a contrast between the following:

$$B_{\text{Elmo}}S[j] \quad (\textit{de dicto}) \quad (7.2)$$

$$(Ex)(x = j \ \& \ B_{\text{Elmo}}S[x]) \quad (\textit{de re}) \quad (7.3)$$

Here $B_{\text{Elmo}} = \textit{Elmo believes that}$.

In spite of the ease at which the distinction (7.2)–(7.3) can be made,

¹⁶ The literature on the *de dicto* vs. *de re* contrast is difficult to survey. One reason why it has commanded so much attention on the part of philosophers is that there is supposed to be a special difficulty about the use of *de re* constructions in the context of modal concepts. Whatever difficulties there may be about the interpretation of quantified modal logic, the *de dicto* vs. *de re* contrast does not contribute to them. For the contrast will later in this paper be seen to be merely an independence phenomenon. The semantics of such phenomena can be mastered by means of GTS without any philosophical problems.

there has been an endless flow of bad papers in philosophical and linguistic journals about the *de re* vs. *de dicto* distinction. For a long time, this cottage industry only seemed to serve to give a bad name to the *de dicto* vs. *de re* distinction and indeed to the so-called philosophical analysis of language generally.

Even though we still consider the literature on the *de dicto* vs. *de re* distinction largely as an exercise in futility, we have come to realize that the logical reconstruction of the distinction exemplified by (7.2)–(7.3) does not close the issue here. The mistake of the linguists and philosophers who have tried to develop horse-and-buggy accounts is that they have disregarded the model-theoretical aspect of the problem and concentrated on the quaint ways in which the distinction is expressed in our “limpid vernacular”, to use Quine’s phrase. But in the latter realm the analysts who are turned on by natural language problems have indeed uncovered a legitimate puzzle. For instance, let us look at (7.3) as a putative explication of the English sentence form (7.1). If one is trying to understand how the English language actually works, there are worrisome questions concerning (7.1) and (7.2)–(7.3). For instance, where does the quantifier in (7.3) come from? There is no trace of it in (7.1). How can English speakers read it into (7.1) so very easily in the absence of any syntactical clues? And how can the two apparently parallel readings (7.2)–(7.3) of one and the same sentence (7.1) be as different in their logical form as they are? Somehow the reconstructions (7.2)–(7.3) do not succeed in bringing English syntax and English semantics together.

It is here that GTS with its recognition of the phenomenon of II comes to our help. To put the true story in a nutshell, the two possible “logical forms” of (7.1) are in reality not as much (7.2) and (7.3) but (7.2) and

$$B_{Elmo}S[j/B_{Elmo}]. \quad (7.4)$$

Here the most important stumbling-blocks to understanding how speakers of English actually handled sentences like (7.1) have been removed. There is no dubious extra quantifier in (7.4) as there is in (7.3). There is a far greater syntactical analogy between (7.2) and (7.4) than there is between (7.2) and (7.3). Even the remaining disanalogy between (7.2) and (7.4), viz. the independence indicator, presents us with an interesting observation rather than with a problem. For the fact that there is no counterpart to “/ B_{Elmo} ” in the English sentence (7.1) is merely a special instance of a wider regularity. With certain qualifications, it is apparently the case that informational independence is not indicated in

English in any syntactical way. Hence, the absence from (7.1) of any counterpart to the independence indicator in (7.4) is merely a case in point. We shall return to the absence of independence indicators from English later in §9 below.

As was seen, the *de dicto* vs. *de re* contrast could be expressed in conventional logical notation, at least in the simplest cases. Can it always be so expressed, or are there instances of the contrast where the independence notation is indispensable? This interesting problem will not be discussed here.¹⁷

8. Other phenomena

Another methodologically interesting phenomenon which can be considered as an example of II is offered to use by the so-called actuality operators.¹⁸ They are illustrated by sentences like the following:

John believes that there are people who persecute him, but some
of them are in reality merely trying to get his autograph. (8.1)

Once I did not believe that I would now be living in Tallahassee. (8.2)

In order to see what the problem with such sentences is we have to note a peculiarity of traditional modal and intensional logics. This peculiarity can once again be formulated most clearly in terms of GTS. There it seems that the semantical games which are the basis of the semantical evaluation of a sentence in a world W_1 always lead us inevitably further and further away from W_1 . Modal and intensional operators mark steps from W_1 to one of its alternative worlds. Furthermore, nested modal and intensional operators mark steps from alternatives to alternatives, etc. No steps in the other direction are possible. The world of unreconstructed modal and intensional logic is thus like the world of Thomas Wolfe: in it, you cannot go back home again.

¹⁷ If the *de re* construction is held not to have existential force, then it will be impossible to express this reading in the conventional notation in all cases. Moreover, the *de re* reading of higher-order expressions certainly does not have any normal existential force.

¹⁸ The most extensive studies of these operators is ESA SAARINEN (1979), pp. 215–327. He also provides references to earlier literature.

The interest of examples like (8.1)–(8.2) lies in the fact that in evaluating them semantically we must make a return trip to a world or, as in (8.2), to a moment of time considered earlier. For instance, in (8.1) we have to consider as it were a member of one of John's belief worlds in which this individual is persecuting John, to take him or her back to the actual world (or, rather, the world in which (8.1) is being evaluated) and to state that in that world considered earlier she or he is merely trying to obtain John's autograph. Thus, the semantics of (8.1) involves a return trip from an alternative world to a world considered earlier. A similar "return trip" was needed to interpret (2.3) above. This is the reason why (2.3) has no formula equivalent to it in the old-fashioned independence-free notation of modal logic.

How can the semantics of such "return trips" be handled? Again, the actual problem history is interesting. There have been two main types of approaches to the problem of actuality operators. On one of them, the metalogical framework is extended. At each stage of the evaluation, we have to consider not only the question of the truth and falsity of a sentence S_2 in the world W_2 which we have reached at that stage of the evaluation process (semantical game), but we must keep in mind (i.e., the rules of evaluation must involve) also the world W_1 from which the evaluation process of the original sentence S_1 started. This kind of theory is sometimes referred to as two-dimensional or multi-dimensional semantics for modal concepts.

In the other main type of approach, the metatheoretical apparatus is left alone. Instead, certain object-language operators are postulated which serve as return trip tickets. Less metaphorically, let one such operator be "DO", where "O" is one of the usual modal or intensional operators. This DO undoes the step prescribed earlier by O; it means we go back to the world in which we were before O was put into operation. DO is known in the trade as a *backwards-looking operator*.

In this kind of notation, (8.1)–(8.2) could be expressed somewhat as follows:

$$B_{\text{John}}(Ex)(x \text{ persecutes John} \ \& \ DB_{\text{John}}(x \text{ is trying to obtain John's autograph})). \quad (8.3)$$

$$\text{Past} \sim B_1 \text{DPast} \text{ (I am living in Tallahassee)}. \quad (8.4)$$

Both these approaches are thoroughly unsatisfactory, for their main concepts do not have any concrete linguistic reality. It is impossible to see

what logico-semantic cash value there is to new dimensions have that are postulated in the former account. And there is little concrete evidence of a widespread presence of backwards-looking operators in English syntax, apart from a handful of expressions like “now”, “actually”, “in fact”, etc.

An account using the idea of informational independence does much better justice to the facts of the case. For instance, in our new logical notation (cf. §3 above), (8.1)–(8.2) could be represented as follows:

$$B_{\text{John}}(Ex)((x \text{ persecutes John}) \ \& \ (T/B_{\text{John}})(x)) \quad (8.5)$$

where $T(x)$ is a shorthand for the complex attribute “is trying to obtain John’s autograph”.

$$\text{Past} \sim B_1((L/\text{Past})(I)) \quad (8.6)$$

where $L(x)$ is a shorthand for “is living in Tallahassee”.

When we thus construe the phenomena that others have tried to deal with by means of many-dimensional modal semantics or by means of backwards-looking operators as independence phenomena, we gain some important advantages. For instance, the fact that backwards-looking steps are not expressed systematically in English syntactically now becomes simply a special case of the more general regularity which we have encountered before and which says that II is not signalled in English syntactically in a uniform way (cf. §9 below).

Also the backwards-looking operator idea is half-way natural only when the step from world to world that is being reversed is the most recent one. When it is an earlier one, semantical rules begin to get very messy. This is seen already from (8.4), whose interpretation is not unproblematic. In contrast, a treatment based on the II idea does not face any such difficulties.

9. Wider perspectives

Instead of trying to discuss yet further particular phenomena, it is in order to register a few of the methodological and other general implications which the discovery of the ubiquity of II has for linguistic semantics.

First, there is the remarkable fact that II is not signalled in English syntactically in any uniform way. There are admittedly certain particular

constructions which require II. For instance, as Jaakko Hintikka has pointed out, quantifiers occurring in different nested prepositional phrases are normally taken to be informationally independent.¹⁹ There are certain other constructions which encourage, perhaps even prescribe, informational independence, but they cover nevertheless only small subclasses of the set of all instances of II and they involve essentially different syntactical indicators of independence. Even though further work might be needed here, it seems fairly clear that II is not indicated in English in any uniform way syntactically.

The absence of syntactical independence indicators from natural languages like English is perhaps not very surprising. The semantical phenomenon of informational independence can affect the interpretation of expressions which belong syntactically to entirely different categories. Any uniform syntactical indicator of II would therefore have to be able to attach itself to constituents that are widely different from each other syntactically, so different that the rest of English grammar will not allow it.

We are dealing here with phenomena that have a great deal of interest for the general theoretical questions of theoretical semantics. The absence of uniform indication of II in English throws a shadow on all syntax-driven treatments of semantics. II is an essential and important feature of the logical form of a sentence in which such independence occurs, in any sense of logical form that logicians and philosophers are likely to countenance. But if this is so, how can the logical form of a sentence be read off from its syntactical generation, as Chomsky seems to suggest?²⁰

However, the constructive implications of our findings are more important than the critical ones. Just because the phenomenon of II can affect the force of so many different kinds of expressions, it is important to be able to recognize the hidden unity behind the apparent differences. What is in common to the *de re* reading of noun phrases, branching quantifiers, and neg-raising? *Prima facie*, nothing. Yet they all have been found to instantiate one and the same phenomenon, which we have also shown how to treat syntactically and semantically in a uniform way. We

¹⁹ See HINTIKKA (1974).

²⁰ In Chomsky, logical form (LF) is a level of the syntactical construction of a sentence determined by the earlier levels of the syntactical generation of that sentence. (Cf. CHOMSKY (1986), especially pp. 66–67.) How, then, can something that is not indicated by the syntax of English at all find its way to the LF?

believe that there are other concrete semantical phenomena in natural languages whose treatment can similarly be made uniform only in terms of GTS.

References

- CARLSON, L. and TER MEULEN, A., 1979, *Informational independence in intensional Contexts*, in: Esa Saarinen *et al.*, eds., *Essays in Honour of Jaakko Hintikka* (D. Reidel, Dordrecht), pp. 61–72.
- CHOMSKY, N., 1986, *Knowledge of Language* (Praeger, New York).
- HENKIN, L., 1961, *Some remarks on infinitely long formulas*, in: *Infinitistic Methods: Proceedings of the Symposium on the Foundations of Mathematics* (Warsaw, 2–9 September 1959) (Pergamon Press, New York), pp. 167–183.
- HINTIKKA, J., 1974, *Quantifiers vs. quantification theory*, *Linguistic Inquiry* 5, pp. 153–177.
- HINTIKKA, J., 1976, *The Semantics of Questions and the Questions of Semantics* (Acta Philosophica Fennica, Vol. 28, No. 4) (Philosophical Society of Finland, Helsinki).
- HINTIKKA, J., 1979, *Quantifiers in natural languages: some logical problems 1*, in: Ilkka Niiniluoto and Esa Saarinen, eds., *Essays on Mathematical and Philosophical Logic* (D. Reidel, Dordrecht), pp. 295–314. (Also in Saarinen 1979.)
- HINTIKKA, J., 1982a, *Questions with outside quantifiers*, in: R. Schneider, K. Tuite and R. Chametzky, eds., *Papers from the Parasession on Nondeclaratives* (Chicago Linguistics Society, Chicago), pp. 83–92.
- HINTIKKA, J., 1982b, *On games, questions, and strange quantifiers*, in: Tom Pauli, ed., *Philosophical Essays Dedicated to Lennart Aqvist* (Philosophical Society and the Department of Philosophy, University of Uppsala, Sweden), pp. 159–169.
- HINTIKKA, J., 1983, *The Game of Language* (D. Reidel, Dordrecht).
- HINTIKKA, J., 1987a, *Game-theoretical semantics as a synthesis of verificationist and truth-conditional meaning theories*, in: Ernest LePore, ed., *New Directions in Semantics* (Academic Press, New York and London), pp. 235–258.
- HINTIKKA, J., 1987b, *Is scope a viable concept in semantics?*, in: ESCOL '86: *Proceedings of the Third Eastern States Conference on Linguistics* (ESCOL, Columbus, OH), pp. 259–270.
- HINTIKKA, J. and KULAS, J., 1985, *Anaphora and Definite Descriptions: Two Applications of Game-Theoretical Semantics* (D. Reidel, Dordrecht).
- JONES, A.J., 1980, *Game Theory* (John Wiley, New York).
- OWEN, G., 1982, *Game Theory* (Academic Press, New York and London).
- PUTNAM, H., 1971, *Philosophy of Logic* (Harper & Row, New York).
- SAARINEN, E., ed., 1979, *Game-Theoretical Semantics* (D. Reidel, Dordrecht).
- SANDU, G. and HINTIKKA, J., (forthcoming), *Neg-transportation as an Independence Phenomenon*.
- VON NEUMANN, J. and MORGENSTERN, O., 1944, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton).

This Page Intentionally Left Blank

HOW NATURAL IS NATURAL LANGUAGE?

JAN KOSTER

*Institute for General Linguistics, University of Groningen, 9712 TG Groningen,
The Netherlands*

1. The two systems and their interface

Since classical antiquity, debates about human language have been concerned with the question of whether it is natural (that is, biologically given) or conventional (determined by social and cultural contingencies). Even if it is granted that language is characterized by both natural and conventional factors, the various classical thinkers in this area stress either the former or the latter aspect. In the middle of the nineteenth century, for instance, a very outspoken biological view was defended by August Schleicher, who considered languages as natural species engaged in a Darwinian struggle for survival. In the late nineteenth century, Schleicher was ridiculed by the American linguist Whitney, who thought it rather self-evident that language is a matter of cultural convention. De Saussure explicitly adopted Whitney's conventional view, which became the dominating view among the various structuralistic schools (although there are some interesting exceptions).¹

The Chomskyan revolution was, among other things, a return to a mainly psychological—and ultimately biological—perspective on language. Unlike Schleicher, Chomsky does not take languages as external Darwinian species. He is rather concerned with language as a form of knowledge underlying the actual creation of external products and their use. This form of knowledge is thought to develop within tight biological

¹ For Schleicher's ideas, see SCHLEICHER (1863, 1869). Whitney's reactions can be found in WHITNEY (1871, 1874). See also KOERNER (1983) for a recent anthology of the debate. De Saussure's position can be found in the *Cours*, for instance in the version edited by Tullio de Mauro (DE SAUSSURE (1916) [1972: p. 26]).

constraints on the human mind. This view has led to a revival of the classical idea of Universal Grammar. In its Chomskyan form, Universal Grammar is the initial scheme specifying attainable human grammars. UG has certain open parameters that are set by actual experience. In important respects, therefore, language learning in this view is fixing the parameters of UG.²

It is, of course, not denied within the Chomskyan perspective that language has many conventional aspects. It is therefore a misinterpretation to say that according to Chomskyan generative grammar language is innate. What is innate is not the end product but rather diverse underlying schemes that are filled in by actual experience.

Clearly, Chomskyan linguistics situates important aspects of language within human biology. The underlying scheme is referred to as *grammar*, particularly as universal grammar. It is also referred to as the innate language faculty or a mental organ, or even a language organ. Whereas the aforementioned view of Schleicher's can be characterized as biological externalism with respect to language, Chomsky's view can be characterized as biological internalism. In both cases, however, important aspects of language fall within a scheme of biological necessity. In both cases there is something like a biology of language.

Chomsky's main argument for his ideas about the innate biological status of major aspects of language is the so-called "poverty of the stimulus" argument. According to this argument, our tacit knowledge of language can be demonstrated to be very intricate and specific, while its complexity cannot be related in any significant way to explicit instruction or any other form of sufficient environmental input. The successful Chomskyan research paradigm of the last 30 years has shown, beyond reasonable doubt in my opinion, that our tacit knowledge of language is indeed quite deep and in most cases fully unconscious and unrelated to environmental contingencies. More generally, Chomsky has characterized his research paradigm as a contribution to Plato's problem: how can we account for the richness and complexity of our knowledge in spite of the limitations of the evidence in learning situations? The proposal of UG, an initial scheme of possible grammars with open parameters, is the tentative answer to Plato's problem in the realm of human language.

Although there is a considerable amount of disagreement on the issues in question, I personally believe from the actual research results of the

² For a recent version of Chomsky's ideas, see CHOMSKY (1986).

generative enterprise that Chomsky's view is correct in fundamental respects. Human language indeed confronts us with a remarkable discrepancy between complexity of knowledge and the poverty of evidence available to the language learner. In fact, the idea of an innate scheme is the only hypothesis that bridges the gap between knowledge and evidence at all. I agree therefore with Chomsky that something important must be innate.

Where I differ from Chomsky, however, is in my interpretation of what is innate. I think that, ultimately, the characterization of the innate scheme as "grammar", universal or otherwise, is misleading. Similarly, I find it misleading to refer to the biological mechanisms underlying the initial scheme as "the language faculty" or "the language organ". In my alternative interpretation, the underlying scheme has nothing, literally nothing, to do with language. Or at least I will conclude that there is no evidence for an intrinsic relation between the initial scheme and what we usually call language. I will conclude, then, that the link between the initial scheme and language as we know it is not established at the level of biological necessity but at the level of human culture, that is, as a matter of human freedom and history.

Somewhat surprisingly, my conclusion seems to follow from the development of some other Chomskyan ideas, particularly from certain ideas about the modularity of mind. According to the modular view of mind, the mind—at some level at least—is not one integrated whole but a structure composed of various more or less autonomous components. Almost all systems of a certain degree of complexity are modular in this sense. Likewise, the language faculty is not a unitary system but something composed of relatively independent subsystems. An important insight formulated by CHOMSKY (1986) is that with respect to language we have to make a distinction between a computational structure and a more or less independent conceptual structure. This distinction between a computational and a conceptual module is, of course, rather crude, because the two main systems are no doubt further divided into sub-modules. For present purposes, however, a twofold division suffices.

The conceptual module includes what is often referred to as knowledge of the world, common sense and beyond. It also includes logical knowledge and knowledge of predicate-argument structure. The computational module concerns the constraints on our actual organization of discrete units, like morphemes and words, into phrases and constraints on the relations between phrases. This can be illustrated with a simple example.

Consider the following sentence:

The father of Mary knows that John loves her (1)

Conceptually speaking, this sentence is about three individuals, *John*, *Mary*, and *the father of Mary*. These individuals are denoted by names and a definite description that involves a kinship term (namely *father*). Furthermore, the sentence involves two relational concepts, “knowing” and “loving”, the former being a relation between an individual and a proposition, the latter being a relation between two individuals.

Someone who knows English knows that *the father* is the one who *knows* here, rather than *Mary* or *John*. Similarly, we interpret this sentence in such a way that *John* loves *Mary* rather than the other way round. Our knowledge of how the various arguments are distributed over the relational concepts is not part of our conceptual knowledge but part of our computational knowledge or syntax. Someone with the same conceptual knowledge but with a different grammar, that is, with a different computational system could interpret the sentence in such a way that *John* is the one who knows something and “the father” is the one who loves *Mary*. I assume with Chomsky that this computational knowledge is independent from the conceptual knowledge and that “knowledge of language” is organized in the two corresponding basic modules.

One of the most fascinating aspects of our language system is that it has the property of discrete infinity. Thanks to recursive rules we are able to produce or understand strings of discrete units of arbitrary length. Other systems of animal communication of infinite range lack this property of discrete infinity. The well-studied communication system of honey bees, for instance, is infinite in the scope of its possible messages, but involves notions of continuity rather than the discreteness of the human system.³

As far as we know, the capacity to handle discrete infinities in connection with conceptual content is unique to humans. This is perhaps one of the most important conclusions we can draw from the attempts of the last twenty years to teach human language to apes. David Premack, Alan and Beatrice Gardner, and several others have tried to teach aspects of human language to chimpanzees by using some other medium than speech. Premack used plastic chips and the Gardners used gestures derived from American Sign Language, a visual language for the deaf. The results of these experiments are very controversial, but in one respect

³ VON FRISCH (1967) is the classical text on bee language.

they seemed to be a complete failure. Apparently, apes are able to learn aspects of our conceptual system (to what degree is a matter of debate), but there is no evidence at all that they are able to learn our computational system with its property of discrete infinity.⁴

Interestingly, the apes also failed to master the concept of discrete infinity in some other crucial aspect. David Premack observed that his apes were not able to master the concept of number. Apes are no doubt able to see the difference between, say, one enemy and five enemies, but counting in the normal recursive sense is completely beyond their capacities to the best of our knowledge.⁵ If Premack's observation is correct, it may be concluded, as Chomsky has persuasively argued, that the development of the capacity to handle discrete infinities is one of the key events in human evolution. The same development, which from one point of view gives the number faculty, gives the capacity to construct an unbounded range of expressions from another point of view. So, according to Chomsky, the capacity of free thought and the uniqueness of human life originated in the combination of the conceptual system, with its more primitive precursors, and the computational system, with its property of discrete infinity. The latter system is perhaps a "spin off" of the development of our neocortex in comparison to the less developed cortex of the ape.

Note, incidentally, that this view of human evolution is modular in the strongest possible sense. The capacity to handle discrete infinities through recursive rules is not considered a unique property of language but rather something that the number system and language have in common. This fact is important for what follows.

I find Chomsky's speculations concerning the origin and nature of human language not only fascinating but also quite plausible. I will assume, therefore, that this is the correct perspective: human language, powerful as it is, originates from the combination of two capacities with an entirely different evolutionary origin.

A very important question which arises at this point involves the nature of the crucial link between the conceptual systems and the computational system of discrete infinity. It seems to me that Chomsky's views entail

⁴ See TERRACE (1979) for bibliographical data and a critical interpretation of the results of Premack and the Gardners.

⁵ PREMACK (1976: p. 262) says the following: "Our attempts to teach counting to the chimpanzees have enjoyed notably little success—so little that even the elementary stages of counting now loom as a far greater challenge than the elementary stages of language."

that this link itself is also a matter of biological necessity, that is, something that originated in our biological evolution and that is established in the brain of the developing child as the result of the human genetic program. It is this view which I would like to challenge.

In a modular system that fulfills certain functions, it is possible that the various modules are radically independent from the resulting complex system. Electromotors, for instance, are components of systems as functionally different as record players, air-conditioning systems, and coffee grinders. In none of these cases does the electromotor have an intrinsic connection with the resulting functional system. It is simply nonsense to say that the purpose of the electromotor is "coffee grinding". The relation between the electromotor and coffee grinding is not intrinsic but entirely accidental. In general, it is a matter of human creativity to what ends certain devices can be used. Similarly, the relation between certain forms of mathematics and their application is not intrinsic but a matter of human resourcefulness. There is, for instance, no intrinsic connection between arithmetic and book-keeping. Book-keeping is a powerful application of arithmetic, to be sure, but we know that arithmetic can just as well be applied to ends other than book-keeping. It is important to bear in mind that even if book-keeping were the only known application of arithmetic, there would still be no intrinsic connection. Mathematical structures have one application, many applications, or no applications at all. In all applications, the connection between structure and function is an arbitrary link established at the level of human culture.

Especially when certain structures have only one known application, it is often tempting to say that the application in question is the function of the structure in question. After Voltaire's Doctor Pangloss in *Candide*, I will call this the panglossian fallacy: the idea that some function is the intrinsic purpose of some structure, because the function in question is the only one fulfilled by the structure in question. It is the erroneous conclusion, attributed to the German philosopher Christian Wolff, that the purpose of the ears is to support the hat, because obviously ears happen to support hats.

No one takes Wolff's conclusion seriously, but curiously, many generative grammarians believe that the purpose of the computational structure they study is its contribution to human language. Otherwise it would not make much sense to refer to the structure in question by the word "grammar" and to refer to the initial state of this "grammar" as "universal grammar" or even "the language faculty".

In my alternative interpretation, the link between the computational

structures we find in human language and the conceptual systems is not intrinsic at all but accidental, just like the link between arithmetic and book-keeping. This means that the crucial link that makes human language so powerful is not a matter of biological necessity in my view, but a matter of human culture. In sum, I agree with Chomsky that human language originates from the linking of two autonomous systems, but as an alternative to the biological linking hypothesis I would like to propose the hypothesis of cultural linking. If the two systems are linked at the level of human culture, our language must have the status of a cultural achievement itself, in spite of its high degree of biological predetermination.

Note that the traditional "poverty of the stimulus" arguments do not favor the hypothesis of biological linking. Suppose, for instance, that arithmetic were innate (counter to fact) and that some hero of human culture had discovered that arithmetic can be applied to commercial concepts. In that case, human children would perhaps learn book-keeping very rapidly and we would no doubt know a lot about book-keeping without explicit instruction. But it would be an error, of course, to conclude from this state of affairs that humans are blessed with an innate book-keeping faculty. What would be innate would be something else, namely arithmetic, which has nothing to do with book-keeping at the relevant biological level.

Similarly, it could be a mistake to conclude from rapid human language learning and "poverty of the stimulus" arguments that there is something like universal grammar or an innate language faculty. So, I do not take issue with the idea that "poverty of the stimulus" arguments show that there are innate structures underlying human language, but I disagree with the usual characterization of these underlying structures. Characterizing them as "universal grammar" is an unwarranted panglossian practice, in my opinion.

Of course, the idea of biological linking of the basic modules could be right after all. But this would be an empirical discovery, either at the level of brain mechanisms or at the level of human evolution. In the latter case, for instance, one would have to show that the basic architecture of the computational module is an *adaptation* to its use in combination with the conceptual modules. There is no evidence for adaptation at all, however. On the contrary, the more we know about the computational module, the less plausible it seems that one module is intrinsically "made" for a functional combination with the other.

First, I will say something more about the nature of the crucial link

between computational and conceptual modules. If the two basic modules are autonomous and radically different in architecture and evolutionary origin, there must be something like what is called an interface in computer systems. Ideas concerning the interface are of course not unrelated to the issue whether the linking is biological or cultural. My hypothesis is that the interface we are looking for is the lexicon and that this fact—if it is a fact—adds to the plausibility of the view that the crucial linking is cultural rather than biological.

Clearly, the words of the lexicon of any language are cultural artefacts not transferred by our genes but by socialization processes. This is true even if there are heavy constraints on possible words, both on their sound shape and on their conceptual content. Lexical items are deliberately created within cultures and must be learned from one generation to the next. Nevertheless, lexical items are indispensable ingredients of human languages as we know them. There are no languages without a dictionary.

If we look at the properties of lexical items, they are indeed the most plausible candidate for the interface we are looking for. Lexical items typically have both a conceptual side and a structural side. They connect the world of our knowledge with the range of possible syntactic structures. The structural aspect of lexical items is so prominent that some even claim that syntactic structures are projected from the lexicon. If my interface hypothesis is correct, this is only apparently so. What the lexicon actually does is to connect two independent systems, namely the computational system and the conceptual system in the sense discussed before.

So far, the lexicon is the only plausible candidate for the interface between the two basic modules. Since the lexicon is a cultural artefact, human language as we know it is itself a cultural artefact resulting from one of the most fruitful inventions in human history, at least as important as the invention of the wheel, printing, or the computer. By using the lexicon as an interface between our computational and conceptual faculties, the full resources of an easily accessible and versatile computational system become available for the expression and combination of conceptual content. In this way, human beings can express an infinite range of thought and use it for a multitude of purposes, including communication.

Our use of the lexicon as an interface between the two systems has some element of free, be it hardly escapable choice. In principle, we could combine our conceptual system with an entirely different computational system. In practice, however, we are not able to get along with other computational systems since these are less versatile or accessible given the limitations on our brain structure.

Before going on, I would like to point out a potential misunderstanding. I have argued so far that language only exists in the world of human culture and history as the result of the invention of the lexicon as an interface between the computational and the conceptual worlds. I have further argued that there is no intrinsic, let alone a biologically necessary link between the computational system and human language, which is such a powerful application of it. One might conclude from this that I am advocating some form of general intelligence status for the computational system rather than a language-specific status. This is not what I am claiming, however. Some aspects of the computational system may have a more general application, particularly the notion of discrete infinity itself as it also appears in the number system. But the system of structure-dependent relations as it appears in natural language computations may be entirely language-specific in the sense that it does not have other known applications. In that case, the computations are trivially language-specific but still arbitrary with respect to the application in which they enter. Likewise, if book-keeping were the only known application of arithmetic the connection would still be arbitrary and non-intrinsic.

In any case, if human language only exists as a cultural artefact thanks to the invention of an interface between two otherwise unrelated types of systems, there is no longer any reason to talk about universal grammar or an innate language faculty, at least not as long as we are talking about the computational system, which has been the main topic of generative research. This conclusion holds no matter to what degree the computational faculty is itself innate in all its richness.

2. The decline of panglossian linguistics

An interesting question is why the panglossian fallacy is so persistent in linguistics, even within the Chomskyan tradition. Chomskyan linguistics, as we saw, is panglossian in that it assumes that the structures it studies exist for the purpose of language. This is clear from its characterization of the computational system as “grammar”, I-language, or even “the language faculty”. Since there is no more reason to assume that the computational structure exists for the purpose of language than there is reason to assume that the hands exist for the purpose of playing the piano, this persistent tradition is curious, to say the least. Why is the computational system not studied as a structure in its own right without inherent meaning or purpose?

One reason is perhaps that the computational structures in question

only appear to us in their application in language. If arithmetic were only known to us disguised as book-keeping, it would not be easy to discover its status as an autonomous structure. Historically speaking, mathematical structures were often known first in some applied form before the more abstract and more general patterns were recognized. Surveying, for instance, is older than the more abstract Euclidean geometry. From that perspective, it is not so surprising that the study of an application, grammar, is older than the study of the underlying abstract pattern.

A second reason is perhaps the long tradition of functionalism in the humanities, something to which I will return.

American linguistics since Bloomfield, including transformational grammar, has not been functionalistic or "teleological" in the European sense. The survival of panglossian elements is therefore probably due to the first reason rather than to the second. The panglossian element was easily preserved by the tendency of early generative grammar to study natural languages in analogy with artificial languages. Artificial languages are sets of well-formed formulas characterized by some syntax. The syntax is functional in the sense that it is deliberately designed as a characterization of the language.

By analogy, natural languages were often seen as sets of sentences that were considered well-formed in some intuitive sense. CHOMSKY (1986) refers to such sets as E-languages. The reconstruction of the concept of "natural language" as E-language does not seem compatible with the cognitive-psychological goals of generative grammar. Since it is realized that natural language is the result of very heterogeneous modules, the idea of an intuitively given set of well-formed sentences has become entirely elusive. One reason is that we can only isolate sets of sentences by making a decision, namely the decision which module we will take into account. A sentence can be acceptable from one point of view and unacceptable from some other point of view. There is no non-arbitrary way to determine which set of sentences is the "real" language. Since the selection of a set of sentences involves decisions, there are no natural E-languages but only artificial ones.

As long as we view the study of natural language as the study of E-languages, there is a considerable tension between this idea and the possibility of obtaining some degree of so-called psychological reality.

Psychological reality is a hotly debated issue, and a large portion of it is just confusion. Thus, I agree with Chomsky that much debate in this area comes down to the question of whether one accepts scientific realism in general. Sentences only have a structure thanks to certain principles of

mental organization. It therefore seems irrational to deny linguists the possibility to learn something about this underlying mental organization by studying the structure of sentences. There are numerous facts we can observe immediately, without special controlled experiments, and there is absolutely no reason to assume that the observations form a less privileged entrance to psychological reality than experiments carried out by experimental psychologists.

Even a realist, however, does not necessarily study psychological reality when he studies reality. Since language has many conventional aspects, there is much that a realist can study in language without doing psychology. Language can also be studied as a product of human culture. This is in fact what most linguists have done most of the time, without becoming non-realists in most cases.

Since language, in my opinion, only exists as a product of human culture, I do not believe that theories are about psychological reality as long as they are about language. If sentences are inventions, resulting from the creative combining of our cognitive modules via the culturally given lexical interface, theories about sentences are not psychological theories. Only theories about the underlying faculties can be psychological in the intended sense.

One such underlying faculty is the computational faculty discovered by generative grammar. At a certain level of abstraction, theories about this faculty are about biologically real structures. As long as we limit ourselves to the biological or psychological level of discourse, these structures have nothing to do with language. It is an error to say that there are such biological structures that entail sets of natural language sentences "together with much else". Since the lexicon is not part of our biologically given nature, the computational and the conceptual modules are not connected at a strictly biological level.

What all this means is that theories such as early generative grammar and its offspring (such as so-called Generalized Phrase Structure Grammar) cannot be psychologically real in principle. The same holds for all other theories that entail sets of sentences (E-languages). This is perhaps why Montague grammarians and other formal semanticists so often have problems with the idea that grammars are psychologically real. Most, if not all, theories of formal semantics assume an explicit syntax of some E-language. This can be interesting or useful for many purposes, but it is something different from the psychological study of the faculties underlying language.

All in all, it seems to me that a non-panglossian approach to the

computational system underlying language throws a new light on the issue of psychological reality.

The most extreme, and often most dogmatic forms of panglossian linguistics are the so-called functional approaches to language. Contrary to what we find in the Bloomfieldian tradition, these approaches have been very influential in Europe, particularly under the influence of the Prague School in the 1920s and 1930s. According to functional linguistics, language must be studied as a "means towards an end", as Roman Jakobson put it.⁶ Ultimately, under these approaches, form must be understood in terms of function.

Prague-style functional linguistics was influenced by Karl Bühler, who saw language as an "organon" (i.e., an instrument).⁷ Functionalism was embedded in the 19th-century Hegelian tradition of the *Geisteswissenschaften*, which played an important role in countries as different as Holland and the Soviet Union. Functionalism with respect to mental phenomena was propagated by Franz Brentano, whose student Husserl had a direct influence on Jakobson and other linguists of the Prague School.⁸ According to the Brentano school, mental phenomena differ from physical phenomena in that they are intentional, that is, goal-oriented. In general such conceptions of the mental go back to the organistic metaphors of Romanticism. The Romantics often opposed the idea of a goal-oriented organ to the mechanistic metaphor of a clockwork, which had dominated the world picture since the scientific revolution of the seventeenth century. Romantic organicism lies at the origin of the wide-spread view, popular in the 19th century and up until the present day, that the living world (and the mental in particular) must be studied in a way different from the one familiar from physics. Intentionality, then, was the hallmark of living nature, while non-living nature was thought to be characterized by causality.⁹

It seems to me that the strictly modular approach to linguistics that I am advocating not only throws a new light on the issue of psychological reality, but also on the issue of functionalism. From the point of view defended here, it makes a crucial difference whether the functionalistic ideas are claimed for language as an integrated system at the level of

⁶ JAKOBSON (1963).

⁷ BÜHLER (1934).

⁸ See for instance HOLENSTEIN (1976: 47ff.).

⁹ See BORING (1950: ch. 17) for an extensive discussion of Brentano's influence. ABRAMS (1953) discusses the Romantic resistance against the mechanical world picture.

human culture or for the underlying modules. No matter how “functional” language use can be shown to be, the conclusion does not carry over to the underlying modules. This particularly holds for the computational system that, according to my hypothesis, is not related to language at all at levels other than human culture (such as the biological or cognitive levels).

More generally, it is clear that physical structures can be related in different ways to their application in human culture. Certain forms of instrumental usage take advantage of entirely accidental properties of physical structures. Primitive man, for instance, made use of the sharp edges of flintstone for the purpose of cutting. It would be absurd in this case to say that flintstone can only be understood functionally, that is, in relation to human butchery. Of course, flintstone is something in its own right. It can be understood physically, the sharp edges can be explained in terms of crystal structure or whatever. There is, in short, no intrinsic connection between flintstone and the accidental use that humans made of its properties.

There are two less accidental ways in which physical structures can be related to some function. The first is conscious design (by human designers), and the second is adaptation by natural selection in the Darwinian sense.

As far as I see, these are the three forms in which physical structures can be functional, by accident, by design, or by adaptation. If this is correct, the question is in which of these three ways is the computational system studied by linguistics related to language. Is our linguistic use of this system a matter of happy accident, or is it a matter of design or biological adaptation?

Design is not a reasonable proposition, because up until recently, no one had ever known of the existence of the computational system. Biological adaptation is a possibility, be it an extremely unlikely one. The reason is that some central features of the computational system have the nature of mathematical essences, which do not seem compatible with the gradualism of Darwinian adaptation.¹⁰ Consider for instance the infinite scope of the system, which it owes to its recursiveness. A system combining discrete units is finite or infinite, there is no gradual Darwinian

¹⁰ Darwinism is usually presented as an anti-essentialistic philosophy (see HULL (1973: ch. V)). In that sense, it is just applied nominalism. Before Darwin, the essentialistic species concept was already undermined by the the semi-nominalism of John Locke (see LOVEJOY (1936: 228ff.)).

transition from finiteness to infinity. Similarly, the computational system shows certain kinds of symmetry. Again, the forms in question seem incompatible with gradualism.

Generally speaking, in spite of the fact that many features of organisms can be seen as adaptations, certain mathematically describable forms have always been exempt from the Darwinian pattern of explanation. What I have in mind are the forms studied in the classical work *Growth and Form* by D'Arcy Thompson and also in the beautiful and equally classic book *Symmetry in Science and Art* by the Soviet scientists Shubnikov and Koptsik.¹¹ It seems to me that what we find in the computational system underlying language is closer to such aspects of nature than to its Darwinian aspects, which can be described in terms of adaptation.

If these conclusions are correct, the only plausible view is that in language we make a happy use of the entirely accidental properties of the computational system. Such properties, then, are in no way intrinsically functional with respect to their use. They are like the sharp edges of flintstone and like the symmetrical patterns we find in organic nature. They are just physical regularities and as such open to mathematical analysis like other aspects of physical nature. Likewise, these structures are best studied like crystals and other manifestations of matter. There is really nothing "human" to these structures, they are just as alien to our aims and needs as quarks and leptons.

Practically everything we know about the structure of the computational system makes functionalism a near absurdity. To what, then, does functionalism thank its appeal, up until the present day? Here we can only speculate, but it seems to me that functionalism is appealing because it pretends to keep language within the teleological space of human intentions. In the alternative (in my opinion right) view, the most fundamental structural aspect underlying language is completely "dehumanized" so to speak. It is not studied in the spirit of the humanities but in the spirit of those fields that concern themselves with non-living matter. Some people find it perhaps frightening to consider such fundamental aspects of the human mind as an ordinary part of physical nature.

There is perhaps yet another aspect to functionalism. If we look at the history of the study of organisms, two fundamentally opposing views can be distinguished. According to one view, organisms have a rich intrinsic nature, determined by their genes. This view is generally accepted for the

¹¹ See D'ARCY THOMPSON (1917) and SHUBNIKOV and KOPTSIK (1974).

physical aspects of organisms. By some curious accident of cultural history, a rich genetically determined nature is often denied for the mental aspects of humans. The human mind has often been considered within a second view of organisms. According to this view, organisms have no intrinsic nature but are plastic and malleable by environmental pressures. This view finds its roots in classical empiricism and has particularly been influential in theories of the human mind. It is heavily ideological and has led to extreme absurdities like American behaviorism. Chomsky has dubbed this theory the "empty organism theory".

To the extent that it does without a heavily constrained physical theory of possible organisms, Darwinism is just the diachronic version of the empty organism theory. It would recognize rich genetic structure at an ontogenetic level, but it would grant extreme plasticity at the phylogenetic level. Without heavy constraints on possible organisms, natural history is just the history of organisms as the result of environmental pressures.

In sum, according to the first theory organisms have a rich intrinsic, physically constrained structure, which is only marginally effected by environmental pressures. According to the opposing view, organisms have no essence, they are plastic entities shaped by their environment.

Functionalism is logically distinct from the empty organism theory, but is nevertheless its natural complement. By denying the autonomy of the structure underlying language, it denies the autonomous status of the structure of the human mind. In that respect, European functionalism is just as environmentalistic as American behaviorism. Ultimately, then, both ideologies are very similar in that they mask or deny the rich intrinsic and autonomous structure of the human mind, while emphasizing the shaping role of the environment.

The empirical discovery of a rich and autonomous structure underlying language does not only refute behaviorism but also functionalism. It is my hope that it also restores something of human dignity, of the idea of a richly structured autonomous individual, whose essence is not determined by oppressive "educational" environments.

References

- ABRAMS, M.H., 1953, *The Mirror and the Lamp: Romantic Theory and the Critical Tradition* (Oxford).
BORING, E.G., 1950, *A History of Experimental Psychology* (Englewood Cliffs, N.J.).
BÜHLER, K., 1934, *Sprachtheorie* (Jena).

- CHOMSKY, N., 1986, *Knowledge of Language: its Nature, Origin, and Use* (New York).
- FRISCH, K. VON, 1967, *The Dance Language and Orientation of Bees*. Translated from German by L.E. Chadwick (Cambridge, Mass.).
- HOLENSTEIN, E., 1976, *Roman Jakobson's Approach to Language: Phenomenological Structuralism*. Translated from German by C. and T. Schelbert (Bloomington).
- HULL, D.L., 1973, *Darwin and his Critics: the Reception of Darwin's Theory of Evolution by the Scientific Community* (Chicago).
- JAKOBSON, R., 1963, *Efforts toward a Means-End Model of Language in Interwar Continental Linguistics*, in R. Jakobson, ed., *Selected Writings II*, (The Hague) 1971, pp. 522-526.
- KOERNER, K., 1983, (ed.), *Linguistics and Evolutionary Theory. Three Essays by August Schleicher, Ernst Haeckel, and Wilhelm Bleek*, with an introduction by J. Peter Maher (Amsterdam).
- LOVEJOY, A.O., 1936, *The Great Chain of Being* (Cambridge, Mass.).
- PREMACK, D., 1976, *Intelligence in Ape and Man* (Hillsdale).
- SAUSSURE, F. DE, 1916, *Cours de Linguistique Générale*, publié par Charles Bally et Albert Sechehaye. Edition critique préparée par Tullio de Mauro (Paris) 1972.
- SCHLEICHER, A., 1863, *Die Darwinsche Theorie und die Sprachwissenschaft* (Weimar).
- SCHLEICHER, A., 1869, *Darwinism tested by the Science of Language*. Translated from German by Alexander V.W. Bikkers (London).
- SHUBNIKOV, A.V. and KOPTSIK, V.A., 1974, *Symmetry in Science and Art*. Translated from Russian by G.D. Archard (New York).
- TERRACE, H.S., 1979, *Nim* (New York).
- THOMPSON, D'ARCY, W., 1917, *On Growth and Form*, J.T. Bonner, ed., (Cambridge) 1977.
- WHITNEY, W.D., 1871, *Strictures on the Views of August Schleicher Respecting the Nature of Language and Kindred Subjects*, *Transactions of the American Philological Association* 2, pp. 35-64.
- WHITNEY, W.D., 1874, *On Darwinism and Language*, *North Amer. Rev.*, 119, pp. 61-88.

13

History of Logic, Methodology
and Philosophy of Science

This Page Intentionally Left Blank

LEIBNIZ AND THE PHILOSOPHICAL ANALYSIS OF SCIENCE

FRANÇOIS DUCHESNEAU

*Département de Philosophie, Université de Montréal, C.P. 6128, Succ. A, Montréal,
Quebec H3C 2J7, Canada*

The ideal of *mathesis universalis* rules over Leibnizian as well as Cartesian methodology. But, at the time Leibniz's system was shaping up, empiricist methodologies were being developed by *experimental philosophers* like Boyle and Hooke, Locke and Newton. The new science had to be built from observation and classification of phenomena; hypotheses should remain subordinate to an inductive ordering of data; and abstract deduction should be rescinded as speculative. Indeed, the mechanistic ideal still afforded patterns for causal explanation: phenomena in their regular sequences were to be explained by reference to modes of extension, figure and motion. But could one get to the true natural processes beyond the inductive and analogical inferences based on experience? Though none of the experimental philosophers questioned the geometrico-mechanistic pattern of intelligibility, there was pervasive skepticism about the possibility of discovering specific "mechanisms" behind the screen of phenomena; and it was doubted whether certain knowledge about real essences could be obtained beyond recourse to analogical models. And so, Boyle and his followers, including Newton, set the corpuscular hypothesis as a provisional foundation for physics. While it provided a perfect "image" of geometrico-mechanistic intelligibility, it also expressed a failure in the attempt to render science certain by reducing its explanatory arguments to *more geometrico* deductions from definitions and axioms. Thus, philosophy would cease serving a foundational function for science; its contribution would become more indirect and "critical"; it would determine the modalities and limits of our access to knowledge depending on the diverse objects of scientific analysis, and help remove cognitive obstacles in the process.

Leibniz's position stands out against that of the experimental philosophers. While he was framing his system and building his dynamics,

he questioned empiricist accounts of the laws of impact that lacked theoretical covering, as those of Wallis, Huygens and Wren. He sketched a series of projects for a scientific encyclopedia, in which more and more room was given to the notion of a *scientia generalis*. Thus he developed a conception of philosophy of science that can be defined as architectonic rather than strictly foundational or critical. This philosophy of science will influence Leibniz's scientific work, especially in the dynamics which comprises plans for theory-building [cf. *Dynamica de potentia* (1689–1690)]. Leibniz seems to have been the first to proceed to analyze what we mean by a theory in science. Starting from his conception of theory-building, he sought to diagnose deficiencies and aporias in the empiricist methodologies [cf. the epistemological developments in the *New Essays concerning Human Understanding* (1704) against Locke and in the correspondence with Clarke, a spokesman for Newton's methodology and philosophy of nature].

Directly relevant to theory-building according to Leibniz is the development of analysis jointly as a method of invention and as a method of demonstration. In the first part of this paper I shall try to sketch the original conception of the analytic method Leibniz developed. In the second part, my purpose will be to underline how Leibnizian analysis aims at dealing in a combinatory fashion with indefinite notions, thus allowing for the potential systematization of truths of fact.

All modern references to the method of analysis stem from the beginning of Book VII in Pappus's *Mathematicae Collectiones* translated into Latin by Commandino.

Resolution (*analysis*) is a process that starts by admitting what is in question and that gets through what follows therefrom, to something which is conceded in composition. For in resolution, we postulate what is in question as resolved; we consider what follows from it; then we do the same for the antecedent to this first step; and we repeat the operation until we fall on something which we already know, or which fits among the principles. And we call this process resolution, as it is a resolution achieved *ex contrario*. In composition (*synthesis*), by reverting the process, we set as given what we admitted at the end of the resolution; we put in order, following nature, a series of antecedents which were consequents in the other process; and once the mutual combination of these is achieved, we get to the end of what is in question; and this mode is called composition.¹

It seems clear that the model for synthesis is taken from geometrical demonstrations. Such demonstrations start from elements : namely definitions and axioms; and they develop the logical implications of these elements into progressively more compounded relations. Hence, synthesis would essentially appear as a process of exposition issuing from constructs

which knowledge of the elements supports and justifies. The tread in demonstration consists in linking terms according to the direct order of conditional implications. Following the inverse path, the analysis of ancient geometers as stated by Descartes in the *Regulae ad directionem ingenii*, formed essentially a process of invention for solving problems.² By supposing the problems solved on figures, one would draw such constructions as afford hypothetical connections with principles qualifying as elements. Thus, one might weave the tread of the direct demonstration by reascending from consequences to premisses for a given conditional implication. Descartes ascribed this method to the Greek geometers as their concealed discovery process. The lack of direct testimony on analytic constructions would be a result of their decision not to divulgate these except to their initiated disciples. The main Cartesian criticism about this analysis of the Ancients focuses on the need to tie the progressive unfolding of hypothetical relations to lines drawn on the figures; as a consequence the understanding is deprived of its autonomy in casting mediating concepts. Descartes's objective in setting forth his own method was to promote the universal validity of analysis both for discovery and demonstration. In fact, it meant freeing analysis from the descriptive route based on the imaginative framing of figures. It also meant developing a method based on the unfolding of simple natures in the conceptual decomposition of complex natures. Descartes wanted also to give rules for exhausting alternative hypotheses so as to get acceptable substitutes for demonstrations hinging on conditional implication, when the decomposition of complex natures can only be relative as is the case with phenomena of nature. Descartes could not escape dealing with validity criteria for these propositions which are used to supply a discursive analytic sequence.

Indeed, the contents of the *Regulae* related to this epistemological task; and one may also consider with some evidence of reason the methodological precepts in the *Discours* as a summary of epistemological conditions that should prevail in identifying valid analytic progressions. But it becomes evident that Descartes refers to a psychological criterion of distinctness of the conceptual ingredients when he attempts to assess the projected relations for reascending towards the principles of a demonstration. Apprehending connections between otherwise distinct concepts forms the process for getting a chain of intuitions on which the analytic demonstration can hinge. In a way, Descartes undervalues the need to test the formal structure of implications in demonstrative inferences. His conception of *illatio* is based on the capacity to link in chains some

distinct perceptions of conceptual relations. The abstract strategy of imposing on concepts formal connections of a general type to warrant the validity of resulting propositions would appear to him a vain artifact of reason. This is precisely what he denounces in synthesis as a mode of exposition. In this regard, his *Responses to the Second Objections* afford clear evidence:

Analysis shows the true way by which a thing was methodically discovered and derived, as it were effect from cause, so that, if the reader care to follow it and give sufficient attention to everything, he understands the matter no less perfectly and makes it as much his own as if he had himself discovered it. But it contains nothing to incite belief in an inattentive or hostile reader; for if the very least thing brought forward escapes his notice, the necessity of the conclusions is lost; and on many matters which, nevertheless, should be specially noted, it often scarcely touches, because they are clear to anyone who gives sufficient attention to them. Synthesis contrariwise employs an opposite procedure, one in which the search goes as it were from effect to cause (though often here the proof itself is from cause to effect to a greater extent than in the former case). It does indeed clearly demonstrate its conclusions, and it employs a long series of definitions, postulates, axioms, theorems and problems, so that if one of the conclusions that follow is denied, it may at once be shown to be contained in what has gone before. Thus the reader, however hostile and obstinate, is compelled to render his assent. Yet this method is not so satisfactory as the other and does not equally well content the eager learner, because it does not show the way in which the matter taught was discovered.³

To illustrate the Cartesian distinction between analysis and synthesis, Brunschvicg used to refer to Pappus's problem which Golius had proposed to Descartes.⁴ This problem, which probably influenced the researches that led to the *Géométrie*, is thus stated:

Given $2n$ straight lines, find the locus of one point such that the product of its distances to n of these straight lines present a determined ratio to the product of its distances to the other n lines.⁵

Following Brunschvicg's interpretation, the effects in the demonstrative order are identified to the lines, the causes to metric relations that determined the position of those lines. Analysis substitutes to the lines metric proportions until the ratio of coordinates for solid loci of n lines is reached. Synthesis may develop various applications of this ratio of coordinates, provided one does not lose sight of the fact that the proof of these applications resides in the analytic apprehension of a sufficient condition for all analogous constructions; and this condition unveils itself only through the analytic process. In a characteristic fashion, Brunschvicg relates the distinction of causes and effects along the analytic order to the distinction which *Regula VI* drew between absolute and relative. This

latter distinction was based on the epistemic dependence ordering the conceptual and therefore real implications of the so-called simple natures. The Cartesian position is particularly significant in this instance:

I call that absolute which contains within itself the pure and simple essence of which we are in quest. Thus the term will be applicable to whatever is considered as being independent, or a cause, or simple, universal, one, equal, straight, and so forth; and the absolute I call the simplest and the easiest of all, so that we can make use of it in the solution of questions. But the relative is that which, while participating in the same nature, or at least sharing in it to some degree which enables us to relate it to the absolute and to deduce it from that by a chain of operations, involves in addition something else in its concept which I call relativity. Examples of this are found in whatever is said to be dependent or an effect, composite, particular, many, unequal, unlike, oblique, etc. These relatives are the further removed from the absolute, in proportion as they contain more elements of relativity subordinate the one to the other. We state in this rule that these should all be distinguished and their correlative connection and natural order so observed, that we may be able by traversing all the intermediate steps to proceed from the most remote to that which is in the highest degree absolute. Herein lies the secret of this whole method, that in all things, we should diligently mark that which is most absolute.⁶

Thus one deals with a conjunction of the requisites for analysis both as a method of discovery and as a method of demonstration. On the one hand, setting a network of conceptual distinctions makes it possible to link through more or less varied and multiple connections the complex terms of problems with sufficient reasons for the order that is involved. On the other hand, the resolute process unfolds by revealing the conceptual ingredients to the conscious self: it does so by sequential perceptions expressing the connected concepts. And so the demonstrative process reascends from effect to cause, by a gradual substitution of conceptual connections to the confusedly perceived manifold in the complex contents of problems: one has to apprehend the cause in the effect as its essential reason. Unfolding the effects from principles is by contrast a blind operation, since it works according to general forms of argumentation, which express but do not explicitly elucidate the effective conditional implications between those conceptual elements which make for the effect to be analyzed.

In contrast with this Cartesian doctrine, the Leibnizian doctrine on analysis purports to replace the intuitive progression by a process of unfolding of the formal structures involved in the concatenation of concepts and propositions. Y. Belaval, among others, judiciously described how Descartes and Leibniz diverge on the methods of invention and demonstration.⁷ He underlined in particular that they refer to different mathematical models: Descartes relied mainly on metric geome-

try, to which he co-ordinated the symbolic techniques of algebra. Leibniz resorts to some mathematics that afford more powerful formal means. The architectonic disposition of Leibnizian mathematics may be shown in a table where everything else gets subordinated to the principles of a logical axiomatic (see Fig. 1).⁸

In the end, as noted by Brunschvicg, Descartes claimed for mathematical analysis its autonomy from the formal implications that logic would reveal; logic seemed to him void of knowledge contents and dispensable; rather, the task at hand would consist in forming chains of intuitions to resolve problems up to principles and instancing the progressive make-up of relations derived from simple natures. As for Leibniz, “at least it is certain that he uses mathematics contrariwise to Descartes, to promote Aristotelian logic; and he can thus bring close together geometers, jurists, schoolmen, disciples of Lulle, physicists, and metaphysicians — and it is also certain that he defends an analytic theory of mathematical argumentation.”⁹

For Leibniz, the validity of mathematics would result from the correspondence one can set between stages of analytical decombination and a series of judgments whose formal validity can be identified for each element in the series. These judgments would express real definitions,

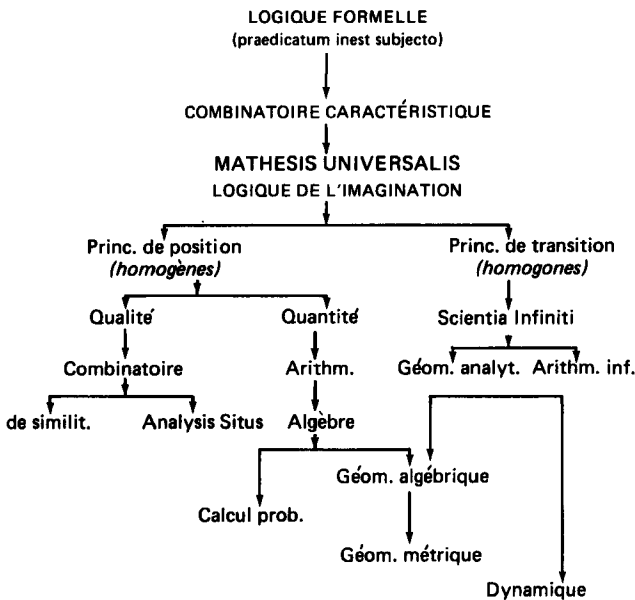


Fig. 1.

setting forth the possibility of the corresponding objects, that is the compatibility of the various conceptual ingredients entering their definitions. While Descartes admitted of intelligible essences only insofar as evidenced in the actual perceiving of ideas, Leibniz reckons such essences only when they get explicated and proven in the shape of real definitions. Thus, ideas cease being taken for objects of the passive intellect, to become signs of implicit judgments that may develop their contents through proper intellectual operations; they would then symbolize with the intelligible order at the foundation of reality.

Indeed, Descartes envisaged also a kind of explication of ideas through analysis, but it meant reascending analytically from relative to absolute in the intuitive apprehending of intelligible conditions for the object to be analyzed. An *inspectio mentis* would circumscribe the essential determining connections that structure the objects and establish their derivation from simple natures. Leibniz focuses on the combination and decombination of concepts that form sufficient marks for reckoning the object as real because of its essential possibility. He feels this will grant him privileged access to the analytic model. When one passes from the empirical to the rational sphere, determining concepts, and propositions in which these concepts would avail, requires that decomposition and composition be continued up to the unveiling of formal compatibility connexions between ingredients, up to primitive concepts, or at least up to concepts whose further resolution is not strictly required for grounding the rational argumentation on the decombination or rather on the achieved part thereof. From this viewpoint, an *inspectio mentis* will never suffice, except in the extreme case of primitive concepts or of primitive truths of reason and fact; furthermore, one needs provide for each stage of the composition–decomposition, the system of requisites ruling over the conditional implications. For demonstration is a *catena definitionum*¹⁰; and a definition is itself a complex of requisites specifying the possibility of the signified object. Analysis consists in expressing the requisites of an adequate definition: this means ordering the requisites in such a way that each adequate expression for the *definiendum* entails its reciprocity with the other possible expressions based on equivalent requisites. A fragment edited by Couturat represents in summary this logical model for analysis operating by definitional resolution.

Resolution is substituting definition in place of defined. Composition is substituting defined in place of definition.

Of the same defined there can be many definitions [This point is illustrated by reference to a combinatoria using letters and integers].

Every reciprocal property can be a definition. A definition is so much the more perfect, when the terms which enter into it are the more capable of resolution. A definition is perfect enough, if once it is explained one cannot doubt whether the defined is possible. If one of the definitions is chosen, all the others will be demonstrated from it, as its properties.

Whichever reciprocal property exhausts the whole nature of the subject; or from whichever reciprocal property, all its properties can be deduced.

A *requisite* is what can enter a definition.¹¹

One should connect these points to Leibniz's argument on the primacy of causal or genetic definitions among real definitions. Such definitions indicate one, if not the sole possible, order of requisites for producing the object; and so, they involve the expression of equipollent series of requisites. Under these circumstances, one can but agree with Belaval in identifying the theory of definition as the "clé de l'analyse" for Leibniz.¹² The perceptive apprehension of intellectual contents may afford but an apparent means for the understanding to reach the essential connections that structure any complex object. Hence, one can fall a victim to illusory resolution. Real resolution is based on the logic of connections justifying demonstrations that abide by the rule. Analysis attempts to find ground for setting forth such connections.

Indeed, analysis is grounded in the requisites of concepts. But, correlatively, these requisites can be viewed as compossible conceptual ingredients that combine to form derivative relations and thus make for the definition of the more complex objects. This is why analysis presents itself indifferently in the so-called analytic and in the synthetic or combinatory mode. For synthesis cannot be a mere exposition technique. If synthesis belongs to the art of judging and inventing, it is because our understanding can apprehend the basic analytic connections in resolving the more complex objects, and because these very objects have to be accounted for by producing the relevant combinative formula.

Many Leibnizian texts witness to the fact that analysis and synthesis integrate under the same architectonic notion of a rational method (*methodus rationis*). For instance, the *De Synthesi et Analysisi universali seu Arte inveniendi et judicandi* (1679)¹³ introduced the distinctions Leibniz had drawn between analysis and synthesis as a corollary to a unified theory of the method of discovery. This theory is based on the *ars combinatoria*: incomplex terms (concepts) are determined and may be set in a systematic order for combinations so that the sequences of complex terms (propositions) can be drawn therefrom. This procedure relates to distinct notions. It is presumed that non-distinct notions, those which are confused or insufficiently determined, can be progressively reduced to

distinctness by various means. Also, the theory of real definitions, based on the normative pattern of genetic or causal definitions, indicates how the *ars combinatoria* would apply to objects of either analytic or synthetic intellection. I shall concern myself here with the role of analysis in this scheme, specially in dealing with notions encompassing an infinite nexus of terms.

Hobbes had asserted that all truths can be demonstrated from definitions. Leibniz did agree with that. But the author of the *De Corpore* had also presumed that definitions result from an arbitrary imposition of verbal signs without rational dependence on the significations to be expressed and without reference to an objective order beyond the determining power of individuals. Leibniz did not admit such a nominalism, which would prevent conceptual analysis from producing real definitions and establishing the notions of possibles *a priori*. Indeed, conceiving possibles does not mean getting access to the effective mode by which objects of experience have been generated. But the doctrine of distinct adequate ideas implies that one can form true judgments on what is contained in the essence considered. Judgments to be true must be consistent with any equipollent series of requisites for the notion. Along this line, analysis should put forth adequate hypotheses that would be equivalent to real definitions of the genetic type. The *ars combinatoria* concerns itself in principle with the full range of possible permutations between equipollent series of requisites and it leaves open the real option among the various combinations which represent the same essential structure. Analysis gains access to one such combination as a sufficient reason for the connections involved in the complex object, but it cannot present any evidence whether such a reduction is ultimate. In some cases, however, significant definitional equivalents can be reconstituted on the basis of the analytic expression that was chosen as *filum inveniendi*, but then, the exhaustion of possible expressions for the requisites can only be ascertained provisionally. This may suffice in most demonstrations, even in geometry where one needs only reduce as much as possible the number of axioms and postulates, pending further analytical demonstrations.

To set up a hypothesis or to explain the method of production is merely to demonstrate the possibility of a thing, and this is useful even though the thing in question often has not been generated in that way. Thus the same ellipse can be thought of either as described in a plane with the aid of two foci and the motion of a thread about them or as a conic or a cylindrical section. Once a hypothesis on a manner of generation is found, one has a real definition from which other definitions can also be derived, and from them those can be selected which best satisfy the other conditions, when a method of actually producing the

thing is sought. Those real definitions are most perfect, furthermore, which are common to all the hypotheses or methods of generation and which involve the proximate cause of a thing, and from which the possibility of the thing is immediately apparent without presupposing any experiment or the demonstration of any further possibilities. In other words, those real definitions are most perfect that resolve the thing into simple primitive notions understood in themselves. Such knowledge I usually call adequate or intuitive, for, if there were any inconsistency, it would appear here at once, since no further resolution can take place.¹⁴

But, defining analysis and synthesis as a twofold expression of the combinative order does not suffice. Leibniz must also take into account the system of resolution-composition that prevails in truths of fact; with the latter, the understanding cannot determine the *raison d'être* of the requisites on the sole basis of logical compatibility. Without entering the mode of composition of contingent truths and the way the principle of sufficient reason rules over them architectonically, it is still appropriate to seek out the methodological processes that illustrate analysis and synthesis as they pertain to truths of fact. Comprised within the primary truths of fact—the assertion of the *cogito* and of the various *cogitata*—the principle which is referred to, states that those concepts are intelligible which would warrant the consensus of phenomena. The analogy here is with the compatibility of definitional requisites for truths of reason. The postulated combinative structure of contingent truths is open for analysis: phenomena can be decomposed with the help of abstract truths of reason serving as tools for transcribing and resolving them: “And so mixed sciences are formed.”¹⁵ In fact, with empirical knowledges, which are more or less formalizable, it is particularly difficult to separate analytical and synthetical modes of discovery and demonstration. The ordering of experiences is at once synthetical, since it results from the comparison of instances, and analytical, since it consists in detecting an order of causal or conditional implications in the manifold of phenomena. Otherwise, the combination of both methodological modes shows up in transcriptions that conform to models drawn from necessary truths; for we deal then with “*praenotiones*”¹⁶ which rule over the ordering of empirical data (synthesis) and help frame such hypotheses as can be translated into conditional implications (analysis). These conditional implications will be adequate if they represent in a consistent and fruitful manner the causal *raisons d'être* for the regular sequences of phenomena (analysis and synthesis combined).

If demonstrations (and inventions) proceed *a priori*, one deals with implications between definitional equivalents, then with systematic derivation and deductive ordering of more or less complex propositions

mediated by formal connections: thus, analysis and synthesis represent the possibility of operative intellection from composite to simple, or from simple to composite. If demonstrations (and inventions) proceed *a posteriori*, one deals with processes that correspond to substitutions of definitional equivalents, but in the form of hypotheses, hinging on extrinsic denominations, symbolic and ultimately inadequate notions, partially adequate and nominal definitions: thus one can carry out some analytic or combinative transcriptions of phenomenal sequences in terms of abstract truths (expressing conditional implications one finds in necessary truths). This may throw light on the intelligibility unfolding *ad infinitum* in truths of fact. It seems that analytic and synthetic patterns of argumentation cannot be dissociated in those hypothetical constructs. Their validity depends on the unfolding of an order proleptically induced and verifiable through successive empirical controls.

The unity of the scientific procedure depends more on the logical features of this disclosure of rationality than on the types of cognition involved. The orientation of the logical processes is that which makes for the difference between analysis and synthesis, but this distinction remains quite relative, since a combinative structure is implied in any object to be known and in any truth to be asserted about it.

Synthesis is achieved when we begin from principles and run through truths in good order, thus discovering certain progressions and setting up tables, or sometimes general formulas, in which the answers to emerging questions can later be discovered. Analysis goes back to the principles in order to solve the given problems only, just as if neither we nor others had discovered anything before.¹⁶

A significant formula is found in a mathematical fragment. It underlines the functional affinity between analysis and synthesis, though the resulting sets of truths may seem disparate:

There are two methods: synthetic i.e. by *ars combinatoria* and analytic. Each can show the origin of invention: this is not therefore the privilege of analysis. The distinction is that the combinatoria starting from the simpler elements exhibits a whole science, or at least a series of theorems and problems, and among these, the one which is sought. Analysis reduces a problem to simpler elements and it does this by a leap, as in algebra, or by intermediate problems in *topica* or reduction. The same distinction holds within combinatoria: for we start either from first elements or from proximate ones.¹⁷

One easily perceives that synthesis sets up axiomatic constructions of propositions: thus for a given domain, one will possess an integrated body of knowledge. The mode *par excellence* for expressing a theory is

therefore synthetic: that ties up with the notion of science as a full set of theorems, as a complete series of implications illuminating the successive problems and their mode of resolution. (This is the case with constructions of the hypothetical type.) On the other hand, Leibniz conceives both analysis and synthesis as occasioning either immediate or deferred constructions and resolutions: the deferred ones rely on series of intermediary results: thus analysis and synthesis can remain provisional while reaching only to mediate steps in the sequence of reasons; and this applies whether one combines derived characters without knowing the primitives, or analyzes by means of imperfect notions as with hypotheses.

The complementarity of analysis and synthesis is also instanced in the plurality of discursive practices that can be interpreted as part of the analytic/synthetic search for order. So, Leibniz notes that syntheses are more general, more theoretical; they serve to discover applications of theories and to frame tables and repertoires¹⁹; they sketch a frame for the encyclopedia as a unified scientific corpus.²⁰ And so, "Those are utterly wrong who think analysis prevails over synthesis, since analysis has been devised to discover perfect synthesis".²¹ Indeed, analysis may easily appear as a means to find the *filum inveniendi* of a synthetic development which will comprise the solution of the problem one started with; analysis may also appear as a device for connecting a problem to be solved with a theoretical corpus.²² In many cases, Leibniz stressed that analysis is not to be used independently of an adventitious synthesis. This is the case when a corpus is already constituted and available: "Analysis is rarely pure, however, for usually, when we search for the means, we come upon contrivances which have already been discovered by others or by ourselves either accidentally or by reason, and which we find stored up as in a table or inventory, either in our own memory or in the accounts of others, and which we now apply for our purpose. But this is synthesis."²³ It is also necessarily the case when one deals with a cryptogram and has to invent a code to decipher it. One needs then rely on some synthetic rules for combining signs in connection with their potential meaning. Also, analysis will not suffice for determining the sum of possible requisites concerning an infinitely complex object, for instance a contingent reality within the system of nature; one has to rely on the consensus of phenomena, which affords a provisional synthesis. References to the cryptogram and to contingent realities combine in this passage:

The analytic method by its own nature may not find the real issue sometimes, the synthetic method always does. As an example where the analytic alone cannot find the

solution take the art of deciphering and other cases where tables have to be set and gone through when we want to know whether a given number is primitive; we also examine possible dividers in an orderly fashion.²⁴

Even if analysis and synthesis can be said to support the same functional procedure, the nature of empirical knowledge, however, puts the emphasis on the analytical method. This method seems to be a necessary condition for all invention, an essential element of art for developing knowledge whenever there is need to cope with an order that unfolds in infinite series. Indeed, this is fundamentally the case with theories to be applied to phenomenal sequences. Analysis generates hypotheses as mainsprings for the unfolding of an order which is at once expressed and dissimulated through complex connections. On the other hand, analysis signalizes itself by not presupposing anything more than that which is strictly required for solving the problem under consideration. To the contrary, the synthetic method, if it could be resorted to in such instances, would give a plurality of possible routes to the solution. They could not all be followed to the point of reaching the specific requisites of the problem. There would be a considerable element of chance if we were to determine the *a priori* demonstration that would contain the sought for solution. Through many occurrences, but specially in the opusculum *Projet et Essais pour arriver à quelque certitude pour finir une bonne partie des disputes et pour avancer l'art d'inventer*, Leibniz presented analysis as containing generically an "art d'inventer admirable".²⁸ He would assimilate analysis with the art of developing demonstration in whichever domain of knowledge, and he would consider it a true continuation of available syntheses, building as it does on elements of knowledge previously organized according to the combinative order. Demonstration is combinative until it needs be transformed into a heuristical tool; analysis is then favored in the search after demonstrative connections. Under these conditions, any system of realities is left open-ended and the inventory may lead to an indefinite number of possible routes:

I find (in the demonstrative connection of propositions) two limits which reason prescribes to us: they are: (1) one needs continue synthesis until it can be changed into analysis, (2) it is useful to continue synthesis until one sees progressions to infinity, (3) when there are nice theorems, specially when they entail practical applications, one should note it also. But the first rule suffices insofar as need goes.²⁶

The third maxim, which does not set a limit to demonstrative sequences as such, concerns mainly theoretical propositions which, issuing from

hypothetical constructs, pertain to the setting data of experiences (adventitious synthesis) in an adequate order. As for the second maxim, it probably applies to demonstrative sequences when they become unable to account deductively for an object involving some indefiniteness, which is indeed the case for contingent realities, whether substances or phenomena. But significantly enough, the first maxim, the only one required for dealing with necessary truths, implies that analysis is aimed at through the various combinative syntheses and forms therefore the proper objective of the *methodus rationis* in furthering invention.

This viewpoint is confirmed both by the Leibnizian description of the analytic method and by the suggested means of generalizing analysis. In the *De arte inveniendi in genere*, Leibniz, after reminding his reader that the analytical method is seldom to be found in a pure state, endeavors to describe it so as to give relief to the formal structure of the arguments it draws together.

First, as often, Leibniz describes models of the method. If one conceives a machinery to be built, joining notions of plain wheels and cog-wheels, the problem is analytical, but solving it depends on preliminary synthetical progressions: that is a case of mixed analysis. But if one has to conceive the intervention of free wheels without axis, analytical necessity yields the notion of a rotation produced by means of a set of teeth. The *De Synthesi et Analysisi universali* presented as more combinative and synthetic the use and application of some acquired invention, for instance using the properties of magnetic needles in devising a compass and as more analytic the conceiving means to meet a duly prescribed technical objective.²⁶ Characteristically enough, the entire methodological procedure is held together by the kingpin of the combinatoria, as a science of forms in general. However, there is an admission that forms or formulas can be comprised in diversely complex wholes, whose sufficient reason analysis may provide, in default of a systematic inventory of all combinations, which remains out of reach. This being said, the essential feature of the analytic method consists in not presuming any requisite, except those which seem strictly indispensable to the projected solution. Such a selective deduction bypasses the synthetic turn of mind, since synthesis must establish all the various formulas in the abstract combination of possibles.

It can evidently happen that the number of requisites be such that the data of the problem cannot be circumscribed in a determinate way. But, then, whatever be the number of solutions left open — and this number can be assigned at least hypothetically — analysis can be restricted to one or a few branches in the combinative differentiation, and for this one or

these few it can attempt to connect the requisites with conditions which have been previously acknowledged. Indeed, determination may never reach such a compatibility of requisites with fully elucidated conditions. Instead of such an *a priori* unfolding, experience may confirm the analytic resolution of requisites by reference to empirically warranted antecedent conditions. This *a posteriori* unfolding reveals at least the latent rationality of the analytically discovered conditions.

Finally, Leibniz suggests that analyzing that which is more particular can be facilitated if we transpose the results of an analytical resolution of that which is more general. This is of major interest when one needs analyze complex phenomena by means of geometrical and mechanical models. In particular, this leaves room for the intervention of determining techniques such as those which the infinitesimal calculus provides. Instancing in this type of determination, Leibniz mentions that the characteristics of the secant make it possible to determine properties of the tangent as a limit in plane geometry.

The following passage in the *De arte inveniendi in genere* summarizes the Leibnizian view on analysis as expressing the *methodus rationis*. And it also underlines the essential idea that analysis represents the heuristical actualization of a real combinatoria, a combinatoria that would encompass phenomena and substances.

In the analytic method, about what is sought, we shall first consider whether it is so determined by the conditions applying to it as object of research, as to be unique; whether it has on the contrary an infinity or an infinite infinity of solutions, or whether it is determined in precise instances. What is sought is a determination of all requisites, or of some only. If of some only, we shall conceive determinations compatible with antecedent determinations, a task which often requires considerable craft. The more determinate we have rendered the thing, the more easily we will get to the solution. It is not always possible to find perfect determinations. Though I have not yet demonstrated something *a priori*, I can see it *a posteriori*, for otherwise all irrational numbers would be rational. When we cannot find more specific considerations, then we survey whether by any chance a more general problem, comprising that one, may not be conceived that would be easier to solve.²⁸

Thus, the formal description of what Leibniz meant by this term reveals the affinity of analysis with the art of framing combinations. The combinative structure of objects to be analyzed is at least anticipated in framing adequate explanatory hypotheses. Analysis, on the other hand, affords the proper means of completing some preliminary syntheses: it sets an account of the requisites for solving problems which depend on previously established and accepted theories. Analysis reflects the aim of a combinative translation for the components of reality whether

phenomena or substances. Despite the indefiniteness of any search for a fully adequate conceptual rendering, and in order to counter infinite regresses in conditions for scientific knowledge, Leibniz construed the analytical method as a tool for systematic hypothesis framing, as a means to achieve consistent and progressive demonstrations whenever infinity prevails in the nexus of terms that would express the order of natural realities.

Notes and references

¹Cf. GILSON, E., 1976, *René Descartes. Discours de la méthode. Texte et commentaire*. 5th Edn. (Vrin, Paris), p. 188.

²Cf. Reg. IV, (1964–1974) in: *Oeuvres de Descartes*, C. Adam and P. Tannery, eds. (Vrin, Paris), X, p. 373–376.

³*The Philosophical Works of Descartes*, 1955, transl. by E. Haldane and G.R.T. Ross (Dover, New York), Vol. 11, pp. 48–49 (A.T. VIII, p. 155–156).

⁴BRUNSCHVICG, L., 1972, *Les étapes de la philosophie mathématique* (Blanchard, Paris), pp. 116–118.

⁵A.T., I, p. 235.

⁶HALDANE and ROSS, I, pp. 15–16 (A.T., X, pp. 381–382).

⁷BELAVAL, Y., 1960, *Leibniz critique de Descartes*, chap. III. La critique des quatre préceptes (Gallimard, Paris), pp. 133–198.

⁸*Ibid.*, p. 137.

⁹*Ibid.*, p. 138.

¹⁰This aspect of Leibniz's doctrine has been considerably underlined by COUTURAT, L., 1969, *La logique de Leibniz* (Olms, Hildesheim), pp. 205–207. Cf. also DUCHESNEAU, F., 1982, *Leibniz et les hypothèses de physique*, *Philosophiques*, 9, pp. 223–238.

¹¹LEIBNIZ, G.W., 1969, *Opuscules et fragments inédits*, in: L. Couturat, ed. (Olms, Hildesheim), p. 258.

¹²BELAVAL, Y., *op. cit.*, p. 194.

¹³LEIBNIZ, G.W., 1965, *Die philosophischen Schriften*, Vol. VII, in: C.J. Gerhardt, ed. (Olms, Hildesheim), pp. 292–298.

¹⁴LEIBNIZ, G.W., 1969, *Philosophical Papers and Letters*, in: L.E. Loemker, ed. (Reidel, Dordrecht), p. 231 (*G*, VII, p. 295).

¹⁵*L*, p. 232 (*G*, VII, p. 296).

¹⁶*G*, VII, p. 296.

¹⁷*L*, p. 232 (*G*, VII, p. 297).

¹⁸*C*, p. 557.

¹⁹*G*, VII, p. 297, *C*, p. 557, p. 159.

²⁰*C*, p. 168.

²¹*C*, p. 159.

²²*G*, VII, p. 297.

²³*L*, p. 233. *G*, VII, p. 297.

²⁴*C*, p. 162.

²⁵*C*, p. 175.

²⁶*C*, p. 180.

²⁷*G*, VII, p. 297.

²⁸*C*, pp. 165–166.

THE LOGICAL IDEAS OF N.A. VASILIEV AND MODERN LOGIC

V.A. SMIRNOV

Academy of Sciences of the USSR, Institute of Philosophy, Moscow, USSR

The original ideas of N.A. Vasiliev, a logician from Kazan, were formulated by him and published in 1910–1913. They did not remain unnoticed. Scientific Russian public showed a great interest for Vasiliev's ideas. Reviews of his articles were published in the *Journal of the Ministry of Public Education*, in the international journal *Logos*, and in the Petersburg newspaper *Rech* ("Speech"). A review on Vasiliev's report about imaginary logic presented by him for the Kazan Scientific Mathematical Society was published in three issues of the newspaper *Kama-Volga's Speech*—a rare event for a logician.

Vasiliev's logical ideas were highly appreciated by N.A. Luzin and Leon Chwistek; his papers are mentioned in the bibliography of A. Church. In post-war years Vasiliev's ideas were investigated by P.V. Kopnin, V.A. Smirnov, G.L. Kline, D.D. Comey, A.I. Maltsev, A. Arruda, V.V. Anosova, and Newton da Costa. V.A. Bajanov from Kazan discovered a series of interesting archive materials by Vasiliev.

N.A. Vasiliev was among the first to put forward the idea of non-classical non-Aristotelian logic. Vasiliev's works actually had a lesser influence on the development of non-classical logics than the works of Lucasiewicz, E. Post and A. Heyting. However, it seems to me that Vasiliev's ideas belong not only to the history of logics, but are topical for the development of contemporary logic as well, and their importance for logic has not yet been fully realized.

At present, the logical ideas of N.A. Vasiliev are attracting the attention of many logicians. He is often considered the forerunner of many-valued logics, intuitionistic logic and para-consistent logic. I consider that such characterizations of Vasiliev are not quite accurate. Undoubtedly, he was one of the first to proclaim and construct non-classical, non-Aristotelean logics. As I shall try to show the type of

non-classical logics proposed by Vasiliev is original, not coinciding with many-valued, intuitionistic or para-consistent logics. This type of logic is to be worked out with the help of modern and powerful logical means.

Nicholas Alexandrovich Vasiliev was born in the city of Kazan, on June 29, 1880, into a highly intellectual family. His grandfather was a famous scientist, and expert in orientology. His father was also a prominent mathematician, an academician, and the editor of the series *New Ideas in Mathematics*. He was greatly interested in his son's ideas and dedicated a book to him. This book was translated into English and the introduction to the book was written by B. Russell.¹

In 1898–1906, N.A. Vasiliev studied medicine in Kazan University and after graduating, served as a country physician till 1904. In 1906, he graduated from another, historical-philosophical department of the same University, defending the thesis *The Question of the Fall of Western Roman Empire in Historiographical Literature in Connection with the Theory of Degradation of Peoples and Human Race*.² In 1907–1910, he was a post-graduate student in Kazan University. In November 1910, he became a Privatdozent of philosophy in Kazan University, in December 1917 a docent, and in October 1918 a full professor in the same university. In the early 1920s he was afflicted by a mental illness, but nevertheless tried now and then to continue his studies. Vasiliev died on December 31st, 1940, in Kazan.

Besides, Vasiliev was a talented poet. He published a book of verses *Longing for Eternity*, and a Russian translation of Verharn's verses. He has also some interesting works in ethics, history of philosophy, and psychology.

But Vasiliev is of interest for us on the basis of his new original ideas on non-classical logics.

The original logical ideas of Vasiliev were formulated and published by him in 1910–1913. He published three articles and the abstract of a Public Lecture. In 1925 he published an abstract for the Fifth International Philosophical Congress. His report about his travels in Germany in 1911–1912 is also very important, and so is the information about one of his reports in January 1911 in the newspaper *Volga-Kama Rech*.

Vasiliev's basic idea is the distinction between two levels in logic. The

¹ I got this information from Professor Eva Zarnecka-Biała (Krakow).

² The full text of the thesis was published in *Izvestiya obschestva arkheologii, istorii i etnografii pri Kazanskom universitete*, 1921, vol. 31, issues 2 and 3.

first or external level is connected with epistemological commitments. It is the logic of truth and falsity. Vasiliev calls it the metalogic. The principles of non-selfcontradiction (an assertion cannot be true and false at the same time) and *tertium non datur* for metalogic (an assertion is true or false) are valid for this logic. According to Vasiliev, metalogic does not vary, it is absolute. In contrast to Vasiliev, it is just this part of logic that is varied, according to Lucasiewicz, Post and Heyting.

The second level in logic depends on ontological (empirical) commitments in relation to cognizable objects. This part of logic can be varied. In Aristotelian logic, the object is not allowed to simultaneously possess and not possess the same properties (law of contradiction).

It is possible to construct other logical systems, the ontological commitments of which are different from those of Aristotelian logic. The empirical part of logic is constructed by N.A. Vasiliev in the form of syllogistic. In his works he gives various systems of non-Aristotelian syllogistic. Imaginary logic (which admits contradiction and law of excluded 4th) and the system of *n*-dimensional logics generalising this imaginary logic are most interesting for us.

Let us axiomatize these and some other logics proposed by Vasiliev.

In his first published article in 1910, *About Particular Statements, Triangles of Oppositions and the Law of Excluded 4th* Vasiliev treats standard particular statements (*J* and *O*) not as categorical, but as problematic.

He introduces three kinds of categorical statements: general positive *ASP*, general negative *ESP*, and particular *TSP* (Vasiliev denotes this statement by "*MSP*"). But, as this term is used as middle, we shall write "*TSP*"). These statements are pairwise inconsistent, and their disjunction is true. I propose the following axioms for this system:

- | | |
|-----------------------------------|--------------------------|
| 1. $AMP \ \& \ ASM \supset \ ASP$ | Barbara |
| 2. $EMP \ \& \ ASM \supset \ ESP$ | Celarent |
| 3. $ESP \supset \ EPS$ | Conversion E |
| 4. $\neg(ASP \ \& \ ESP)$ | |
| 5. $\neg(ASP \ \& \ TSP)$ | Law of the contradiction |
| 6. $\neg(ESP \ \& \ TSP)$ | |
| 7. $ASP \vee \ ESP \vee \ TSP$ | Law of the excluded 4th |

Let us denote this system C1V. This system is definitionally equivalent to C1, formulated in terms of *A*, *E*, *J*, *O*. (Its axioms are: Barbara, Celarent, Conversion *E*, $ASP \supset \ JSP$, $JSP \equiv \neg ESP$, $OSP \equiv \neg ASP$.) The

system $C2V = C1V + 8$. $ESP \vee ASS$ is of interest for us. It is definitionally equivalent to $C2 = C1V + JSP \supset ASS$.

Following Vasiliev, we proposed above an axiomatics of syllogistics with the operator T "only some S are P ". This system proved to be definitionally equivalent to the system of syllogistics with standard operators. However, the ideas of Vasiliev were much deeper than a simple reformulation of standard syllogistics. The decisive factor in understanding the structure of categorical judgements in Vasiliev is the division of judgements into factual ones and judgements about notions. Judgements on facts are judgements stating the results of observations or experiments. Judgements on notions are judgements expressing laws, or nomological statements in modern terminology. According to Vasiliev, a factual judgement expresses something existing, "*was ist*"; a notional judgement, on the other hand, expresses something significant, a rule, a law, something "*was gilt*". A factual judgement cannot express a law; it cannot be valid for a reality that is beyond observation.

On the other hand, a notional judgement cannot express existence. It expresses a law, a connection between existents, but not existence itself. Such a classification of judgements is well known from the history of philosophy. What is new in Vasiliev's work is probably the logical approach to the problem. Vasiliev was the first to propose a logical construction for nomological statements.

The starting point of Vasiliev's analysis of factual judgements are judgements of the form " a is P ", where " a " is a proper name. From singular factual statements it is possible to construct complex factual statements. It is possible to construct group judgements " a, \dots, a_n are P ". If class S consists of a_1, \dots, a_n , then we can pass from a group of judgements to a general factual judgement. "All my friends were faithful to me in need"—is an example of a general factual judgement. Analogously, we can obtain a general negative factual judgement. However, the generalization of a group judgement may follow another (a different) course. From singular judgements we can pass over to numerical judgements, for instance, "Three students of this group got unsatisfactory marks". Finally, it is possible to formulate indefinite numerical judgements—"several S are P ". According to Vasiliev, the particular judgements JSP and OSP of standard syllogistics are indefinite numerical judgements.

For factual judgements A, E, J, O , we have the usual square of opposition, and the system of syllogistics in standard form holds for them.

Vasiliev emphasizes that a judgement “only some S are P ” is not a simple sum of the judgements “some S are P ” and “some S are not P ”, for both of these judgements are factual judgements.

Let us consider this problem more attentively. In his work *About Particular Judgement* Vasiliev gives two interpretations of particular judgements. One of them is disjunctive, the other accidental.

Let us begin with the first interpretation. A disjunctive judgement: “all S are P_1 or $P_2 \dots$ or P_n ”, is understood by Vasiliev as a judgement about the distribution of S among the members of disjunction.

We can write a disjunctive judgement “All S are P_1 or \dots or P_n ” in the terms of predicate calculus—in the following way:

$$\exists x(Sx \ \& \ P_1x) \ \& \ \dots \ \& \ \exists x(Sx \ \& \ P_nx) \ \& \ \forall x(Sx \supset P_1x \vee \dots \vee P_nx) .$$

Then the exclusive particular judgement “All S are P or non- P ” can be written in its disjunctive interpretation in the following way:

$$\exists x(Sx \ \& \ Px) \ \& \ \exists x(Sx \ \& \ \neg P_nx) .$$

An exclusive particular judgement is the judgement about the distribution of objects S between P and non- P . A particular judgement in disjunctive interpretation, according to Vasiliev, is a general judgement, on a par with general, affirmative and general negative ones. A general affirmative judgement says that class S is included in a class P . A general negative one says that S and P are mutually disjoint. A particular judgement speaks about the class S as a whole, about the distribution of the objects S between P and non- P . The system C2V, proposed by me, satisfies the disjunctive interpretation of the judgement TSP . TSP is equivalent to the conjunction of JSP and OSP .

But Vasiliev proposed also another interpretation of exclusive particular judgements—the accidental one. Besides general factual judgements, there are general affirmative and general negative judgements, expressing laws, necessary connections.

Let us adopt the following notation:

$A^{\square}SP$ for “all S are necessarily P ”, $E^{\square}SP$ for “all S are not necessarily P ”, $T^{\vee}SP$ for “all S may be P and may be not P ”. How must we interpret general affirmative judgements of nomological type, “all S are necessarily P ”?

Here it seems natural to give the following translations of the above judgements into the modal predicate calculus S5 π :

$$\begin{aligned} \exists xSx \ \& \ \forall x(Sx \supset \Box Px) & \text{ for } A^\Box SP, \\ \forall x(Sx \supset \Box \neg Px) & \text{ for } E^\Box SP. \end{aligned}$$

However, according to this translation, a general negative judgement turns unconvertible because $\forall x(Px \supset \Box \neg Sx)$ is not deducible from $\forall x(Sx \supset \Box \neg Px)$ in any acceptable modal system. I propose the following translations of these judgements into one-place modal predicate calculus $S5\pi$:

$$\begin{aligned} \theta(A^\Box SP) &= \exists x \Box Sx \ \& \ \forall x \Box (Sx \supset \Box Px) \\ \theta(E^\Box SP) &= \forall x \Box (Sx \supset \Box \neg Px). \end{aligned}$$

It is easy to prove that the θ -translation of $E^\Box SP$ is convertible in $S5\pi$. What meaning do exclusive particular judgements acquire in the accidental interpretation? According to this interpretation, P is accidental in relation to S . A deeper approach would be to consider that every object S may be P and non- P .

However, this interpretation does not coincide with the law of excluded 4th given by Vasiliev. In his first paper Vasiliev would come back to this problem, and at a deeper level, in his later articles. However, we shall now limit ourselves to the first article. Let us denote the exclusive particular judgement under accidental interpretation: $T^\nabla SP$. The translation of this formula into the modal predicate calculus is not difficult to find, if we accept the law of excluded 4th

$$A^\Box SP \vee E^\Box SP \vee T^\nabla SP$$

and the pairwise incompatibility of A^\Box , E^\Box and T^∇ .

In this case

$$T^\nabla SP \equiv \neg A^\Box SP \ \& \ \neg E^\Box SP.$$

Hence

$$\theta(T^\nabla SP) = \exists x \Diamond (Sx \ \& \ \Diamond Px) \ \& \ \exists x \Diamond (Sx \ \& \ \Diamond \neg Px).$$

Thus, we have got a modal interpretation of nomological and accidental judgements—in the terminology of Vasiliev, of notional judgements. On this basis we can understand Vasiliev's assertion that an accidental

judgement is not equivalent to the conjunction of affirmative and negative indefinite numeral judgements (*J* and *O*). *TSP* is deduced from *JSP* and *OSP*.

Vasiliev writes about this most convincingly: an accidental judgement is justified (proved) when there are two factual indefinite-numeral judgements that differ only qualitatively. Several *S* are *P*. Several *S* are not *P*. This is given to us by experiment. Hence we may conclude: "All that is included in the conception *S* is *P* or non-*P*" (Vasiliev, About particular judgement, p. 23).

An accidental judgement is a judgement about a rule, but it follows from factual judgements. It is easy to see that translation of $T^{\forall}SP$ follows from translations *JSP* and *OSP*. The conversion does not hold.

The System C2V is adequate for the disjunctive reading of *TSP* and for the assertoric factual reading of *ASP* and *ESP*. Will C2V be correct for the accidental reading of *TSP* and the apodeictic reading of *ASP* and *ESP*? The answer is negative. All axioms C2V, except 8, are true in the latter interpretation. However, the translation of axiom 8 is not true.

Let us consider axiom 8. $E^{\square}SP \vee A^{\square}SS$.

We can decompose it into two:

$$8_1 E^{\square}SS \supset E^{\square}SP$$

$$8_2 E^{\square}SS \vee A^{\square}SS.$$

The first axiom asserts that an empty class is always included in any class. The translation of this axiom is provable in one-place modal predicate calculus $S5\pi$. However, the translation of $8_2 E^{\square}SS \vee A^{\square}SS$, that is,

$$\forall x \square (Sx \supset \square \neg Sx) \vee \exists x \square Sx \ \& \ \forall x \square (Sx \supset \square Sx)$$

is not provable in $S5\pi$.

Let us denote by *CVA* (Vasiliev's syllogistic in the accidental form) the system that we get from C2V by replacing the axiom 8. $E^{\square}SP \vee A^{\square}SS$ by axiom $8_1. E^{\square}SS \supset E^{\square}SP$. There is a theorem: If α is provable in *CVA*, then $\theta(\alpha)$ is provable in $S5\pi$. Two interesting questions arise now: is the proposed translation an immersion operation? Is it possible to extend *CVA* to a system which is definitionally equivalent to $S5\pi$?

In the end of his article "Particular judgements", Vasiliev remarks that the division of judgements into factual and notional ones covers not only

non-singular, but singular judgements as well. Not only judgements “ a is P ” and “ a is not P ”—are singular, but also judgements “ a is necessarily P ”, “ a cannot be P ”, and “ a can be P and can be not P ”—are also singular. This topic becomes the object of a thorough investigation in the article *Imaginary (non-Aristotelian) Logic* and in less formal, and more general article *Logic and Metalogic*.

Vasiliev wants to construct an imaginary, non-Aristotelian logic. This logic is founded on the assumption of contradiction in conceivable objects. Vasiliev does not admit any contradiction in the real world of existing objects.

Investigations by N.J. Lobachevsky on the problem of Non-Euclidean Geometry greatly influenced logical works of Vasiliev, especially the article *Imaginary (non-Aristotelian) Logic*. Vasiliev wants to construct an Imaginary, non-Aristotelian Logic. This logic admits contradictions in imaginary things. Vasiliev assumes that there are no contradictions in the world of existing things. Lobachevsky had constructed non-Euclidean Geometry on the axiom opposite to the axiom of parallels. Analogically an imaginary logic can also be constructed that denies the ontological law of contradiction, according to which no property can both belong and not belong to the object.

First of all Vasiliev gives a thorough analysis of the negation of the usual, Aristotelian logic. He proceeds from the presumption that an experiment gives us directly only singular positive judgements. We have no sense organs for the observation of the absence of properties of objects.

Negative sentences are the result of deduction. Suppose that I say: “This book is not red”. I have no way of observing directly the absence of red colour, but I see that this book is yellow. Knowing that the object cannot simultaneously be both red and yellow, I deduce from this observation and from my knowledge of the incompatibility of yellow and red colours, that this book is not red. The very incompatibility of red and yellow is, of course, an ontological characteristic of our world. In another world such incompatibility may not exist.

Now let us assume that the subject has the capacity of observing not only the presence but also the absence of a property. In this case the negative singular judgement is based on experience, just as the positive. There is a symmetry between them. In this case, the possibility of getting simultaneously knowledge about the absence and presence of a certain property depends on some external conditions. In Aristotelian logic a negative sentence coincides with the assertion of the falsity of the positive

sentence, and is essentially a complex sentence. In imaginary logic, a singular negative sentence has an independent character and does not coincide with the assertion of the falsity of the positive sentence. This is a very deep thought. It gives us a possibility of introducing a contradictory description of a state of affairs, and gives us possibilities for constructing relevant and paraconsistent logics.

Instead of two types of singular sentences—positive and negative—which may be compatible, it is possible to introduce three types of singular, atomic sentences—positive, negative and indifferent—which are pairwise inconsistent. In imaginary logic, an atomic positive sentence is “ a is P ”, an atomic negative sentence is “ a is not P ”, and an atomic indifferent sentence is “ a is and is not P ”. “Is and is not” is a separate independent relation.

There is no mysticism about this. Vasiliev proposes two interpretations of atomic sentences. According to the first, the three types of atomic sentences are interpreted in the following way: a is necessarily P ; a is not necessarily P ; a is accidentally P . The second interpretation regards the above-mentioned types of atomic sentences as factual in relation to an imaginary world with actual contradictions.

I propose a topological interpretation of atomic positive, negative and indifferent sentences. Let P^0 be the interior of P , P^+ the closure of P , P^1 the complement of P , and let P^x be the frontier of P . Then the atomic sentences will be $a \in P^0$ (positive), $a \in P^{10}$ (negative), $a \in P^x$ (indifferent).

These three sentences are pairwise incompatible and their disjunction is true.

Let us note that we could deal with two atomic sentences, a positive and a negative one, by interpreting positive sentences as $a \in P^+$ and negative sentences as $a \in P^{1+}$. In this case positive and negative sentences may both be true, i.e. $\neg(a \in P^+ \ \& \ a \in P^{1+})$ is not true, and the law *tertium non datur* $a \in P^+ \vee a \in P^{1+}$ is true. We shall proceed on the basis of the former approach, that is, we shall proceed from three kinds of pairwise incompatible atomic sentences.

On the basis of three kinds of atomic sentences, Vasiliev constructs 7 types of complex sentences: three types of general sentences and four types of particular (accidental) sentences. The general sentences are:

“All S are P ”—general assertive sentence, $A_{\square}SP$

“No S is P ”—general negative sentence, $E_{\square}SP$

“All S are and are not P ”—general indifferent sentence, $A_{\vee}SP$

Particular, accidental sentences:

- (1) "Some S are P , and all others are non- P "
accidental positive-negative, $T_n^p SP$
- (2) "Some S are P , and all others are P and non- P "
accidental positive-indifferent, $T_i^p SP$
- (3) "Some S are non- P , and all others are P and non- P "
accidental negative-indifferent, $T_i^n SP$
- (4) "Some S are P , some S are non- P and all others are
 P and non- P "
accidental positive-negative-indifferent, $T_d SP$.

I propose the following topological interpretation of these sentences:

$$\psi(A_{\square} SP) = \exists x S^0 x \ \& \ \forall x (S^0 x \supset P^0 x)$$

$$\psi(E_{\square} SP) = \forall x (S^0 x \supset P^{10} x)$$

$$\psi(A_{\nabla} SP) = \exists x S^0 x \ \& \ \forall x (S^0 x \supset P^x x)$$

$$\psi(T_n^p SP) = \exists x (S^0 x \ \& \ P^0 x) \ \& \ \exists x (S^0 x \ \& \ P^{10} x) \ \& \ \forall x (S^0 x \supset P^0 x \vee P^{10} x)$$

$$\psi(T_i^p SP) = \exists x (S^0 x \ \& \ P^{10} x) \ \& \ \exists x (S^0 x \ \& \ P^x x) \ \& \ \forall x (S^0 x \supset P^0 x \vee P^x x)$$

$$\psi(T_i^n SP) = \exists x (S^0 x \ \& \ P^{10} x) \ \& \ \exists x (S^0 x \ \& \ P^x x) \ \& \ \forall x (S^0 x \supset P^{10} x \vee P^x x)$$

$$\psi(T_d SP) = \exists x (S^0 x \ \& \ P^0 x) \ \& \ \exists x (S^0 x \ \& \ P^{10} x) \ \& \ \exists x (S^0 x \ \& \ P^x x).$$

Let us note that this interpretation does not give the conversion of general negative sentences E_{\square} . This coincides with the assertions of Vasiliev. In his article *Logic and Metalogic* he writes: "The conversion of a positive judgement is conducted in the same way as in our (that is standard, Aristotelian—V.S.) logic, however, the conversion of negative and indifferent judgements is not possible."³

Seven principal (basic) categorical judgements of Vasiliev's imaginary logic compose the basis, that is they pairwise are incompatible ($7(A_{\square} SP \ \& \ E_{\square} SP)$, $7(A_{\square} SP \ \& \ T_n^p SP)$ and so on) and their disjunction is true:

$$A_{\square} SP \vee E_{\square} SP \vee A_{\nabla} SP \vee T_n^p SP \vee T_i^p SP \vee T_i^n SP \vee T_d SP$$

³ N.A. Vasiliev, *Logic and Metalogic*. Logos, pp. 67–68.

There exist the following modi—the first figure:

$$A_{\square}SM \ \& \ A_{\square}MP \supset A_{\square}SP$$

$$A_{\square}SM \ \& \ E_{\square}MP \supset E_{\square}SP$$

$$A_{\square}SM \ \& \ A_{\nabla}MP \supset A_{\nabla}SP$$

Even in the system C2V it is possible to introduce, as has been shown above, indefinite-particular judgements *JSP* and *OSP*. In his imaginary logic Vasiliev also introduced special indefinite particular judgements. He calls them excluding or preparatory kinds. Following the ideas of Vasiliev, we shall introduce three indefinite particular (excluding, preparatory) kinds of sentences:

“In the least some *S* are necessarily *P*”— $J_{\square}SP$

“In the least some *S* are necessarily not *P*”— $O_{\square}SP$

“In the least some *S* are and not are *P*”— $J_{\nabla}SP$

These sentences may be naturally defined—in general and accidental terms—in the following way:

$$J_{\square}SP =_{df} A_{\square}SP \vee T_n^p SP \vee T_i^p SP \vee T_d SP$$

$$O_{\square}SP =_{df} E_{\square}SP \vee T_n^p SP \vee T_i^n SP \vee T_d SP$$

$$J_{\nabla}SP =_{df} A_{\nabla}SP \vee T_i^p SP \vee T_i^n SP \vee T_d SP$$

It is easy to see that the translations of these judgements into predicate calculus with topological operators are the following:

$$\psi(J_{\square}SP) = \exists x(S^0x \ \& \ P^0x)$$

$$\psi(O_{\square}SP) = \exists xS^0x \ \supset \ \exists x(S^0x \ \& \ P^{10}x)$$

$$\psi(J_{\nabla}SP) = \exists x(S^0x \ \& \ P^x x)$$

In their turn the operators T_n^p , T_i^p , T_i^n and T_d may be defined in the terms of operators J_{\square} , O_{\square} , J_{∇} :

$$T_n^p SP =_{df} J_{\square}SP \ \& \ O_{\square}SP \ \& \ \neg J_{\nabla}SP$$

$$T_i^p SP =_{df} J_{\square}SP \ \& \ J_{\nabla}SP \ \& \ \neg O_{\square}SP$$

$$T_i^n SP =_{df} O_{\square} SP \& J_{\nabla} SP \& \neg J_{\square} SP$$

$$T_d SP =_{df} J_{\square} SP \& O_{\square} SP \& J_{\nabla} SP$$

An axiomatics for the imaginary logic of Vasiliev will be proposed below.

Having constructed imaginary logic with three kinds of atomic sentences (positive, negative and indifferent), Vasiliev puts forward a program of constructing logics with N -kinds of atomic sentences and with the law of excluded $n + 1$.

Let us reconstruct Vasiliev's ideas about polydimensional logics. Let us begin with monodimensional logic. There are only positive singular sentences.

On the basis of these sentences we construct positive sentences ASP and JSP . Let us formulate now the positive system of syllogistic C2V1:

1. $ASM \& AMP \supset ASP$
2. $JSM \& AMP \supset JSP$
3. $ASP \supset JSP$
4. $JSP \supset JPS$
5. $JSP \supset ASS$

System C2V1 can be extended to the system C2V1D, which is definitionally equivalent to lower semilattice with zero. With this aim in view we add an operation of intersection and D for an empty class. To the axioms 1–5 we add the following:

6. $ASM \& AMP \supset AM(S \cap P)$
7. $J(S \cap P)M \supset JSP$
8. $JSP \supset A(S \cap P)S$
9. $JSP \supset A(S \cap P)P$
10. $\neg J\phi\phi$

For the proof of the definitional equivalence of these two systems we add the definition

$$S \subseteq P \equiv \neg ASS \vee ASP$$

to C2V1D, and the definitions

$$ASP \equiv \neg(S \subseteq \phi) \& S \subseteq P$$

$$JSP \equiv \neg(S \cap P \subseteq \phi)$$

to the semilattice.

V.M. Popov proved that C2V1 may be extended to a system definitionally equivalent to quasi-Boolean algebra.

We shall get a standard Aristotelian logic if we add axioms of pairwise incompatibility A and E , A and O , and E and J :

$$\neg(ASP \& ESP)$$

$$\neg(ASP \& OSP)$$

$$\neg(ESP \& JSP)$$

and *tertium non datur*

$$ASP \vee OSP$$

$$ESP \vee JSP$$

to system C2V1.

The system C2V2, obtained in this way, is equivalent to the system C2 given above.

In axiomatizing an n -dimensional logic we meet difficulties, because for logics with dimensions more than 2, a general negative sentence is not conversible.

Let us give a scheme of axioms for n -dimensional ($n \geq 1$). The language has the operators A and J with indices. Let

$$A_1SP = A_{\square}SP \quad J_1SP = J_{\square}SP$$

$$A_2SP = E_{\square}SP \quad J_2SP = O_{\square}SP$$

$$A_3SP = A_{\nabla}SP \quad J_3SP = J_{\nabla}SP$$

Let the axioms of the syllogistic C2Vn be the following:

1. $A_1SM \& A_iMP \supset A_iSP$
2. $J_1SM \& A_iMP \supset J_iSP$
3. $A_1SM \& J_iSP \supset J_iMP$
4. $A_iSP \supset J_iSP$

- $$\begin{array}{l}
 5. \quad \neg(A_i SP \ \& \ J_j SP), \quad i \neq j \\
 6_1. \quad A_1 SP \vee \dots \vee J_i SP \vee \dots \vee J_n SP \\
 6i. \quad J_1 SP \vee \dots \vee A_i SP \vee \dots \vee J_n SP \\
 6n. \quad J_1 SP \vee \dots \vee A_n SP \\
 7. \quad J_1 SP \supset J_1 PS \\
 8. \quad J_1 SP \supset A_1 SS
 \end{array}
 \left. \vphantom{\begin{array}{l} 5. \\ 6_1. \\ 6i. \\ 6n. \\ 7. \\ 8. \end{array}} \right\} \text{for } n \geq 2$$

If $n = 1$, we shall have a monodimensional syllogistic

If $n = 2$, we shall have an Aristotelian syllogistic

If $n = 3$, we get the imaginary syllogistic.

The idea of n -dimensional logics was conveyed to Vasiliev by poly-dimensional geometry. I believe that the idea of polydimensional logic is not equivalent to the idea of many-valued logics. I consider that this idea opens new prospects in the development of non-classical logics.

References

Works of N.A. Vasiliev:

- 1 Longing for Eternity (Russian), Kazan, 1904.
- 2 Verharn, E. (Russian). In Verharn, E., Mad Villages. Kazan, 1907. Translation by N. Vasiliev.
- 3 Report on the first year of studies (from January 1, 1907 till January 1, 1908). Library of Kazan University, Manuscript N 5669.
- 4 Dreams of the Old House. In: Creativity (Russian), Kazan, 1909, pp. 100–107.
- 5 Darwin's Significance in Philosophy (Russian). Kamsko-Voljskaya Rech. 30–1, 1909.
- 6 About Gogol (Russian). Kamsko-Voljskaya Rech. 20th and 25th March, 1909.
- 7 Swinburne's Poetry. (Russian). Vestnik of Europe. St. Petersburg, August 1909, Vol. 4, pp. 507–523.
- 8 Swinburne (Russian). In: Creativity, pp. 137–138.
- 9 Third International Philosophical Congress. Heidelberg (Russian). Journal of the Ministry of People's Education, pp. 53–86.
- 10 About Particular Judgement, the Triangle of Oppositions and the Law of Excluded Fourth (Russian). Uchenye Zapiski of Kazan University, Vol. 77, Book 10, 1910, October, pp. 1–47.
- 11 Imaginary Logic (Russian). Kazan, Society of People's Universities, 1911, p. 6.
- 12 Imaginary non-Aristotelian Logic (Russian). Ministry of People's Education. New Series. Vol. 40, 1912, August, pp. 207–246.
- 13 Report of a Private Docent, Philosophy, Chair of Kazan University, N.A. Vasiliev on the Course of his Studies. 1–VII, 1911—1—VII, 1912. The Library of Kazan University, Manuscript N 6270.
- 14 Logic and Metalogic (Russian). Logos, 1912–1913, Vols. 1–2, pp. 53–81.
- 15 Review of the book "Encyclopaedia der philosophischen Wissenschaften in Verbindung mit W. Windelband herausgegeben von A. Ruge. Logos 1–2, 1912–1913, pp. 387–389.

- 16 Review of the book Paulhan, F., *La logique de la contradiction*. Logos 3–4, 1913, pp. 3663–3365 (Russian).
- 17 Review of the Book of Prof. Geyzer, J., *Lehrbuch der allgemeinen Psychologie*. Logos 1–2, 1912–1913, p. 392 (Russian).
- 18 Review of the book by Henri Poincaré, *Dernières Pensées*. Paris, 1913. Ernest Flammarion—editor. Logos 3–4, 1913 (Russian).
- 19 Logical and Historical Method in Ethics (On Ethical Systems of L.N. Tolstoy and V.S. Solovjev) (Russian). In the book: A collection of articles in the honour of D.A. Korsakov. Kazan, 1914, pp. 449–457.
- 20 Lectures on Psychology. Kazan, 1914, p. 226 (Russian).
- 21 Program on psychology. Kazan University, 1915, p. 6 (Russian).
- 22 The Problem of the Fall of Western Roman in Historiographical Literature in Connection with the Theory of Degradation of Peoples and Humanity (Russian). Notes of Archaeology, History and Ethnography. Kazan University, 1921, vol. XXXI, Numbers 2–3.
- 23 Imaginary (non-Aristotelian) Logic. *Atti del V Congresso Internazionale di filosofia*. Napoli, 1925.

Works on N.A. Vasiliev:

- 1 ANOSOVA, V.V., 1984, *N.A. Vasiliev's logical ideas and paraconsistent logical systems*, Thesis for the degree of Candidate of Philosophy (Moscow) (Russian).
- 2 ANOSOVA, V.V., 1985, *Non-classical negation in "imaginary" logic of N.A. Vasiliev*, Materials of IV Soviet-Finnish logical colloquium "Intensional Logics and Logical Structure of Theories" (Tbilisi) (Russian).
- 3 ANOSOVA, V.V., 1982, *Paraconsistent logics and N.A. Vasiliev's logical ideas*, in *Philosophical of Modal and Intensional Logics* (Moscow) (Russian).
- 4 ANOSOVA, V.V., 1982, *Interconnections between N.A. Vasiliev's logical ideas and many-valued logics*, in *Modal and intensional logics: VIII All-Union Conference on Logics and Methodology of Science* (Moscow) (Russian).
- 5 ARRUDA, A.I., 1900, *N.A. Vasiliev: a forerunner of paraconsistent logic*, in VII International Congress of Logic, Methodology and Philosophy of Science (Salzburg) vol. 6, pp. 14–17.
- 6 ARRUDA, A.I., 1984, *N.A. Vasiliev: A forerunner of paraconsistent logic*, in *Philosophia Naturalis*, vol. 21, pp. 472–491.
- 7 ARRUDA, A.I., 1979, *A survey of paraconsistent logic*, in A.I. Arruda, R. Chuaqui and N.C.A. da Costa, eds., *Proceedings of Fourth Latin-American Symposium on Mathematical Logic* (North-Holland).
- 8 ARRUDA, A.I., 1977, *On the imaginary Logic of N.A. Vasiliev*, in: Arruda, da Costa and Chuaqui, eds., *Non-Classical Logics, Model Theory and Computability* (North-Holland).
- 9 BAZHANOV, V.A., 1986, *Formation and development of N.A. Vasiliev's logical ideas*, *Filosofskiye Nauki* (Moscow) 3, pp. 74–82 (Russian).
- 10 BAZHANOV, V.A., 1987, *The making and development of N.A. Vasiliev logical ideas*, VIII International Congress on Logic, Methodology and Philosophy of Science. Abstracts, vol. 3 (Moscow) pp. 45–47.
- 11 GESSEN, S.I., 1910, *On N.A. Vasiliev's paper "On particular judgements, the triangle of oppositions, and the law of excluded 4th"* (Kazan, 1910), *Rech* (Petersburg), October 11 (Russian).
- 12 GESSEN, S.I., 1910, *Review of N.A. Vasiliev's "On particular judgements, the triangle of oppositions, and the law of excluded 4th"*, *Logos*, book 2, pp. 287–288 (Russian).

- 13 GRENIEWSKI, H., 1958, *Refleksje na marginesie Wykładów z dziejów Tadeusza Kotarbińskiego*, *Studia filozoficzne*, Dwumiesięcznik, 3(6), pp. 176–177.
- 14 IANOVSKIY, V.L., 1911, *Review of N.A. Vasiliev's paper "Non-Euclidean geometry and non-Aristotelean logics"*, presented at a meeting of the Physical-mathematical Society, Kamsko-Volzhszkaya Rech, Kazan, January 16, 19, 22, 25 (Russian).
- 15 JAMMER, M., 1974, *The Philosophy of Quantum Mechanics* (John Wiley & Sons, New York) pp. 342–343.
- 16 KLINE, G., 1965, *N.A. Vasiliev and development of many-valued logic*, in: Anna-Teresa Tymieniecka, ed., *Contribution to logic and methodology in honor of J.M. Bocheński* (North-Holland, Amsterdam).
- 17 KONDAKOV, N.I., 1975, *A Logical Dictionary* (Moscow) (Russian).
- 18 KOPNIN, P.V., 1951, *On a classification of judgements*, *Journal of Tomsk State University* 16 (Russian).
- 19 KOPNIN, P.V., 1950, *On N.A. Vasiliev's logical views: From the history of Russian Logics*. Works of Tomsk State University, vol. 112 (Russian).
- 20 KOPNIN, P.V., 1973, *On N.A. Vasiliev's logical views*, in: P.V. Kopnin, ed., *Dialectics, Logics, Science* (Moscow) (Russian).
- 21 KORCIK, A., 1955, *Przyczynek do historii klasycznej teorii—opozycji zdań asertorycznych*, *Roczniki filozoficzne*, No. 4.
- 22 ŁADOSZ, J., 1961, *Wielowartościowe Rachunki zdań a rozwój logiki*, *Warszawa*, s. 37, 38, 175.
- 23 LAPSHIN, I.N., 1917, *Gnoseological Investigations*, issue 1 (Petersburgh) (Russian).
- 24 LOSSKY, N.O., 1922, *Logics, part 1: Judgement. Concept* (Nauka i Shkola Publishers, Petersburg) pp. 153–158.
- 25 MALCEV, A.I., 1976, *Selected Works*, vol. 1 (Moscow).
- 26 MOROZOV, V.V., 1973, *A retrospect view*, in: *Selected Issues of Algebra and Logics* (Novosibirsk) (Russian).
- 27 POVARNIN, S.I., 1921, *Introduction to Logic* (Petersburgh) (Russian).
- 28 SMIRNOV, V.A., 1987, *Axiomatization of N.A. Vasiliev's logical systems*, in *Modern Logics and Methodology of Science* (Moscow) pp. 143–151 (Russian).
- 29 SMIRNOV, V.A. and STYAZHKIN, N.I., 1960, *Vasiliev*, in: *Philosophical Encyclopedia*, vol. 1 (Moscow) (Russian).
- 30 SMIRNOV, V.A., 1987, *Logical Methods Analysis of Scientific Knowledge* (Moscow) pp. 161–169 (Russian).
- 31 SMIRNOV, V.A., 1962, *N.A. Vasiliev's logical views*, in *Essays in the History of logics in Russia* (Moscow) (Russian).
- 32 SMIRNOV, K.A., 1911, *Review of N.A. Vasiliev's paper "On particular judgements, the triangle of oppositions, and the law of excluded 4th"*, *Journal of the Ministry of Public Education* (The New Series, part 32, March, 1911) (Petersburgh) pp. 144–154 (Russian).
- 33 SMIRNOV, V.A., 1987, *Logical ideas of N.A. Vasiliev and modern logic*, VIII International Congress of Logic, Methodology and Philosophy of Science, vol. 5, part 3 (Moscow) pp. 86–89.
- 34 SMIRNOV, V.A., 1986, *Modality de re and Vasiliev's imaginary logics*, *Logique et Analyse*, vol. 114.
- 35 COMEY, D.D., 1965, *Review of V.A. Smirnov 1962*, *The Journal of Symbolic Logic*, vol. 30.

PHENOMENALISM, RELATIVITY AND ATOMS: REHABILITATING ERNST MACH'S PHILOSOPHY OF SCIENCE

GEREON WOLTERS

Institute of Philosophy, University of Konstanz, Konstanz, FRG

1. Introduction

With the exception of German speaking countries no other country in the world responded so quickly, so extensively, and so favorably to Ernst Mach's work as did pre-revolutionary Russia. The reason for this was not only the fact that even then the Russian intelligentsia had an astonishing knowledge of foreign languages. For when the Russian editions of Mach's *Mechanics* and *Knowledge and Error* were published in 1909, all of Mach's major works in philosophy and the history and philosophy of science were available in Russian with the exception of the still untranslated *Principles of the Theory of Heat*. Likewise, in 1909 a book appeared here in Moscow published under the pseudonym of Vladimir Ilyin. It seems as if this book was later to put an end to an unbiased reception of Mach's works in the Soviet Union and other socialist countries. This comes as a surprise because Vladimir Ilyin was neither a physicist nor a philosopher, but a lawyer. Moreover, the work on the book, including the printing, took him less than a year. Essentially, the book consists of a tiresome series of quotes within the framework of an epistemological polemic. This polemic is directed, on the one hand, against Ernst Mach (1838–1916) and Richard Avenarius (1843–1896) and, on the other, and this is the actual aim of the book, against some of the Russian followers of these two philosophers. In my opinion, no one would read this book any more if the author had not had another pseudonym: Vladimir Ilyitsh Lenin (1870–1924). The title of the book is: *Materializm i ėmpiriokriti-cizm. Kritičeskie zametki ob odnoj reakcionnoj filosofii* (Materialism and

Empirio-criticism. Critical Remarks on a Reactionary Philosophy).¹ Unfortunately, the fact that Lenin was a political theoretician, revolutionary fighter and statesman of eminent historical significance has led some to the erroneous conclusion that he had something important to say about epistemology and the philosophy of science. This is not the case. At the same time, Lenin's book seems to have sealed the fate of Mach's writings on philosophy and philosophy of science in socialist countries. For only a few academics have had the intellectual autonomy as well as the courage to deal without bias with an author who had been branded at the highest political level as a reactionary and "fideist", i.e. a supporter of religious ideas.² Lenin himself, if he had lived to see it, would probably have been amazed at the unexpected progress of his book from an incidental political polemic to a highly reputed classic of epistemology. In this paper, however, I do not wish to examine Lenin's misunderstanding or lack of understanding of Mach's phenomenalism. What I am more concerned with is rehabilitating Mach's philosophy of science. With this in mind I would first like to show (in Section 2) that Mach's phenomenalism is not an integral part of his philosophy of science and that as a result Mach's philosophy of science can be acceptable to those who reject his phenomenalism. In the two subsequent sections, I will discuss the view

¹ By the way Mach learned rather quickly of the existence of this book which first appeared in German in 1929. Friedrich Adler had already written him on July 23, 1909: "In Russia the 'battle over Mach' goes on. The enclosed copy of a letter from Kautsky that I had published in 'Der Kampf' [Der Kampf 2 1908/09: p. 451f.] might possibly interest you. Unfortunately, it won't help much. Plechanoff already explained that Kautsky understands nothing about the question and now the dispute is beginning to be pursued energetically in the other party faction. Lenin (Bogdanoff's party comrade) has published a 440 page book with the title: *Materialism and Empirio-criticism. Critical Remarks on a Reactionary Philosophy* in which you, Avenarius, and all of your supporters are abused and all arguments that someone can find who doesn't understand the subject are put together very prettily. Lenin did not concern himself with philosophy before and now has spent a year on it in order to understand why people have gone "crazy". He was certainly very industrious, and in a short time really worked his way through all the literature, but naturally did not have the time to think his way through. He actually considers the elements to be a deceitful trick. It is really unfortunate that I only know the book piecemeal from what my wife has translated for me, but it is enough to see that one will probably not find any serious argument in what he has written."

² More than twenty years ago (1966) F. Herneck had already demonstrated how ridiculous these charges are biographically with respect to Mach. Mach was one of the very few university professors in German speaking countries who stood up for the rights of the working class and its party: the social democratic party. In addition Mach was a pugnacious anti-cleric and probably an atheist. Herneck's article, including a "postscript", is easily obtainable in HERNECK (1986: 109-155).

that Mach's philosophy of science failed because Mach rejected two decisive theories of modern physics, the theory of relativity and the atomic theory. With respect to the theory of relativity (Section 3) it can be proved that the relevant texts published under Mach's name were forgeries. As far as the theory of atoms (Section 4) is concerned, I would like to show that in the last years of his life Mach probably even came to believe in the existence of atoms. Section 5 contains a concluding summary.

2. Mach's phenomenalism and his philosophy of science

Mach was a phenomenalist. As far as this paper is concerned this means that he held only a slightly modified version of a view first advocated systematically by George Berkeley (1685–1753), the view that all of our knowledge is constituted by the data of consciousness. Phenomenalism, however, is not the basis of Mach's philosophy of science. On the contrary, Mach's philosophy of science has a *practical foundation*.³ It is precisely this indissoluble relation of theory to praxis that made Mach's conception so attractive for socialist theoreticians. It was as attractive for bolsheviks like A.A. Bogdanov (1873–1928) as it was for social democrats like Friedrich Adler (1879–1906), the physicist and later general secretary of the International Union of Socialist Worker's Parties.

How should we conceive of this relation of praxis to theory? As Mach described it in the introduction to his *Mechanics*, science developed (1) *historically* out of man's interaction with nature. It seems to him "natural to assume that the instinctive collection of experiences preceded the scientific classification of them".⁴ In its beginnings, then, science presents itself as the *theoretical* continuation of elementary, *practical* orientations in life. Viewed sociologically, the origins of science are closely connected with the development of a class of scientists: "The transition to the classified, scientific knowledge and apprehension of facts first becomes possible with the development of special classes and professions that make the satisfaction of particular social needs their objective in life. A

³ On the question of the practical foundation of science in general, see Jürgen Mittelstraß's instructive study (MITTELSTRASS 1972).

⁴ Ernst MACH (1933: *Mechanik* 4, Engl. *Mechanics*, 5). On several occasions we have corrected the English translations of Mach's works.

class of this sort occupies itself with special kinds of natural processes" (MACH 1933: *ibid*). The class of scientists inevitably develops special forms of interaction: "The individuals in this class change; old members drop out and new ones come in. Thus arises a need of imparting to those who are newly come in the stock of experience and knowledge already possessed" (MACH 1933: *Mechanik* 4f, *Mechanics* 5). Out of this practical, social necessity of being initiated into the forms of knowledge and action in the class of scientists Mach reconstructs (2) *systematically* the fundamental determination of all theory in the philosophy of science. Here it is necessary to "describe to him [i.e. the new member] the phenomena in some way" (MACH 1933: *Mechanik* 5, *Mechanics* 6f.). What can actually be described? "That only can be described, and conceptually represented, which is uniform and conformable to law [. . .]. Thence is imposed the task of everywhere seeking out in the natural phenomena those elements that are the same, and that amid all multiplicity are ever present" (*ibid.*). When it is successful "this ability leads to a *comprehensive, compact, consistent, and facile conception of the facts*" (*ibid.*). The foundation and goal of all science is precisely this, *stating the facts*. For Mach everything that goes beyond stating the facts is "metaphysics". Metaphysics, and naturally the view that some truths are only accessible through "fideism" or occultism, has no place in science, although metaphysical ideas can also have a heuristic value.⁵ Stating the facts as the deepest internal objective of science implies all the other principles in Mach's philosophy of science. These principles are concerned with the question of how facts can be stated most reliably. Valid principles are, among others, the following: (1) a *principle of reality*, according to which only what is in principle observable can be considered a fact, (2) a *principle of economy*, according to which the broadest possible area of facts must be represented with the least possible conceptual means⁶ (the principle of economy is to be understood in the sense of simplicity and range of application in current philosophy of science), and (3) characterization of scientific process as (a) adaptation of ideas to facts and (b) adaptation of ideas to one another. Adapting ideas to facts means observation. Adapting ideas to one another means theory. Theory,

⁵ With respect to Robert Mayer, one of the co-discoverers of the energy principle (see *Mechanik* 481, *Mechanics* 608).

⁶ Mach's principle of economy is reminiscent of W. Whewell's 'consilience of inductions', which consists in subsuming theories in areas which are heterogenous on first sight (e.g. electricity, magnetism, optics) under one new general theory (e.g. electromagnetic theory) (see BURTS 1973: pp. 53–85).

in other words, always transcends mere observation. Theoretical propositions have always to take into consideration other, already established, theoretical propositions.⁷

Up to now I have not mentioned Mach's phenomenalism in connection with this description of Mach's conception of science. The reason for this is quite simple: up to this point phenomenalism does not play any role in Mach's conception. Mach's philosophy of science does not imply a theory of knowledge, but presents itself as a consistent reconstruction of historical praxis. Phenomenalism only comes into play in the next stage: "If this point of view is kept firmly in mind [i.e. science restricted to the representation of the factual] in that wide field of investigation which includes the physical and the psychical, we obtain, as our first and most obvious step, the conception of the sensations as the common elements of all possible physical and psychical experiences, which merely consist in the different kinds of ways in which these elements are combined, or in their dependence on one another" (MACH 1923: *Analyse* X, *Analysis* XI). This means that phenomenalism only becomes relevant for the philosophy of science in those special cases where in one and the same investigation domains are viewed together which are completely separate from one another according to a naively realistic understanding. Mach's prime example is the investigation of the body-mind relationship. Here cognizing is involved in examining itself. For this reason, the naively realistic dualism of knowing subject and an external world existing independently of it is in Mach's view an obstacle to the correct stating of the facts. Mach has nothing against naive realism. Every individual, including Ernst Mach, is an epistemological realist. For realism is a result of the evolution of organisms. "It has arisen in the course of immeasurable time without the intentional assistance of man. It is a product of nature, and it is preserved by nature. Everything that philosophy has accomplished [...] is, as compared with it, but an insignificant and ephemeral product of art." Correspondingly, for Mach naive realism "has a claim of highest consideration" (MACH 1923: *Analyse* 30, *Analysis* 31) Now it is often the case that a structure or a function that has evolved over a period of time can become unsuitable or even harmful under changing conditions. Our genetic repertoire of behavioral traits, for example, evolved during the Pleistocene or ice age when the homonoids lived in small hordes. Today this genetic repertoire represents a liability (perhaps the greatest) in securing the survival of mankind in the face of

⁷ One should notice here the resemblance to Pierre Duhem's "theoretical holism".

nuclear and ecological dangers. Though cognitive abilities have evolved over time they do not secure any guarantee of correctness. For example, on the one hand, we *visually perceive* space as non-homogeneous and anisotropic and this makes good evolutionary sense. *Geometry and physics*, on the other hand, tell us that space is locally Euclidean, and from this it follows that within the realm of visual perception space is also “in reality” homogeneous and isotropic. For Mach, correspondingly, realism as a product of human evolutionary development has its limits. These limits are reached whenever areas are investigated that played no role during evolution, for example, when cognizing man begins to *reflect* on his own cognizing. Outside of these areas, it is naturally expedient and even unavoidable that we follow our realistically oriented, evolutionary disposition: “The human being who at the moment only strives after a practical goal (such a one is often enough also an academic, the working physicist, indeed even the philosopher who at the moment does not want to be critical)⁸ does not necessarily have to give up his instinctively acquired, natural [i.e. naively realistic] conception of the world that automatically guides him in his actions. Like the common man, he can speak of ‘things’ that he wants to grasp, put on the scale, he can speak of ‘objects’ he wants to examine and behave accordingly”, that is, according to his naively realistic mode of perceiving, thinking, and acting (MACH 1923: *Analyse* 303, not contained in *Analysis*). In short, phenomenism is not the basis of Mach’s methodology. It is rather the attempt through an “epistemological turn” (MACH 1923: *Analyse* X, *Analysis*, XI) to come to more adequate solutions within a particular domain of knowledge: “A whole series of troublesome pseudo-problems disappears at once. The aim of this book is not to put forward any system of philosophy, or any comprehensive theory of the universe” (*ibid.*) by presenting a phenomenistic approach to psychophysics. Mach, of course, is convinced of the methodological expediency as well as the epistemological acceptability of his phenomenism. Like every position in philosophy or in a special science phenomenism is also subject to constant criticism and control: “This fundamental [phenomenistic] view (without any pretension to being a philosophy for all eternity) can at present be adhered to in all fields of experience; it is consequently the one that accommodates itself with the least expenditure of energy, that is, more economically than any other, to the present *temporary collective* state of

⁸ The word “critical” is used here in the sense of “critique” of knowledge. See Kant’s *Critique of Pure Reason*.

knowledge. Furthermore, in the consciousness of its purely economical function, this fundamental [phenomenalistic] view is eminently tolerant [. . .]. It is also ever ready, upon subsequent extensions of the field of experience, to give way before a better conception" (MACH 1923: *Analyse* 26, *Analysis* 32).

I hope I have been able to show that Mach's phenomenalism is not the basis of his philosophy of science, but only has a special methodological function for the scientific investigation of heterogeneous areas like psychophysics. My aim was not to justify Mach's phenomenalism, but only to determine his position within the philosophy of science. For Mach, phenomenalism certainly possessed more than just this methodological function. Phenomenalism expresses Mach's philosophical world view. However, this world view does not play any role in his philosophy of science. Despite Mach's sympathy for socialism, he held a partial, i.e. politically biased theory of knowledge and philosophy of science to be just as non-sensical as a partial physics. On July 11th, 1906 he wrote to Friedrich Adler: "You are probably well aware that our unforgettable A. Menger planned to write a social democratic theory of knowledge. To me it was always comical that a social democrat, where something practical and not pure knowledge is involved, should think differently from an individual with an attachment to some party."⁹

3. Mach and the theory of relativity

Ever since Max Planck (in 1908) called Mach a "false prophet" who shall be known by the fruits of his labor (PLANCK 1965: p. 51), opponents of Mach's philosophy of science have referred to the "infertility" of his teachings as a means of supporting their views. Although this is normally a thoroughly acceptable procedure, there is another witness we can call upon whose importance should not be underestimated: Albert Einstein. The evidence that Mach played a highly important role particularly in Einstein's discovery of the general theory of relativity ranges from a postcard Einstein wrote to Mach in 1909 to a statement Einstein made in an interview with the American historian of science, I. Bernard Cohen, 46 years later, two weeks before Einstein's death in April, 1955. Einstein

⁹ The original is in the Adler Archive of the Association for the History of the Worker's Movement in Vienna (Adler-Archiv des Vereins für die Geschichte der Arbeiterbewegung).

closed his postcard to Mach with “your devoted student, Albert Einstein”.¹⁰ Cohen reports: “Although Einstein did not agree with the radical position adopted by Mach, he told me he admired Mach’s writings, which had a great influence on him” (COHEN 1955: p. 72). Mach’s “radical position” undoubtedly refers to his phenomenalism. What remains and what Einstein emphasized again and again throughout his life was Mach’s eminent influence on his philosophy of science.¹¹ Mach’s contemporary opponents rely on a number of different strategies in order to cover over this simple fact. One strategy is to recall that Einstein once called Mach a deplorable philosopher.¹² This restricts Mach’s indisputable influence on Einstein to mechanics and thus physics, and excludes in that way any influence in philosophy of science. Such an approach, however, does not inspire much confidence. For the fundamental idea underlying Mach’s mechanics, namely that “everything that happens in the world is to be explained by the interaction of material bodies”, can admittedly already be found “in its essential form in Spinoza” (KUZNETSOV 1977: p. 345). This can mean nothing other than that the fundamental idea underlying Mach’s conception of mechanics is the sort of *philosophical* idea that constitutes the foundation of all science. Moreover, one should also take into consideration that Einstein’s characterization of Mach as a “deplorable philosopher” only appears once in his writings. He made this comment at a meeting of the “Société française de philosophie” in Paris on April 6th, 1922. The probable reason for Einstein’s comment is that he had just learned that in the foreword to the posthumously published *Optik* Mach had rather bluntly rejected the theory of relativity as well as the attempt to praise him as one of its forerunners (MACH 1921: *Optik* VIIIIf, *Optics* VIIIf). Einstein must have taken this personally since Mach’s alleged rejection in the *Optics* preface refers vaguely to “correspondence”. In his letters to Mach, Einstein had always stressed Mach’s influence and in his responses as well as probably during Einstein’s visit in Vienna in September 1910, Mach never contradicted this.¹³ On the contrary, even as late as new year

¹⁰ A copy of Einstein’s letter to Mach, fully edited for the first time by F. Herneck in 1966, is given by HERNECK (1986: p. 134) as well as by WOLTERS (1987: p. 150).

¹¹ I have tried to provide a detailed and comprehensive description of this influence in Chap. 1 of WOLTERS (1987).

¹² See e.g. KUZNETSOV (1977: p. 345). In differentiating between the “deplorable philosopher” Mach who should be rejected and the highly valued “mechanic”, however, Kuznetsov adopts Einstein’s own and equally unfounded use of terms in this case.

¹³ For a more detailed account see here and in the following WOLTERS (1987), *passim*.

1913/14 Mach, who was by then 76 years old, signalled to Einstein his “friendly interest” in the first and still unsatisfactory version of the field equations of gravitation, that is, the general theory of relativity. The statements in the *Optik* preface must have made Einstein feel that he was being maliciously deceived. Later he accounted for the *Optik* preface in a different way. He traced it back to Mach’s age and illness, which must have prevented him from understanding the theory of relativity and its closeness to his own philosophy of science. At first glance this explanation is not implausible. Mach was already 67 in 1905, the year Einstein published an article “On the Electrodynamics of Moving Bodies” that was to become the foundation for the special theory of relativity. For the last seven years, ever since 1898, however, Mach had been paralyzed on his right side and suffered severely from a number of related complications. In 1913, when Einstein’s first work with the field equations of gravitation appeared, Mach was 75 and his state of health considerably worse than in 1905. However, as plausible as Einstein’s explanation may be, it has the disadvantage of being incorrect. This is because I believe I am capable of proving that the *Optik* preface was *almost certainly forged*.¹⁴ The forger is Ernst Mach’s oldest son, *the physician Dr. Ludwig Mach*.

The correctness of the forgery thesis implies particularly that Ernst Mach did not at all reject the theory of relativity. For the *Optik* preface is the only document that can be used to support his alleged rejection.¹⁵ It can furthermore be shown that Mach followed the development of relativity theory with *critical sympathy*. He did that from the perspective of a man who set something in motion in the development of which he was not able to take part.

Mach seems to have taken a fundamental interest in relativity theory only in 1909 after having read the published version of Hermann Minkowski’s famous speech, “Raum und Zeit”. As a result he added footnotes in three publications in which he indicated the continuity of relativity theory with his own views. This continuity can be demonstrated in detail. In May 1914 in his correspondence with the “Machist” (Lenin)

¹⁴ The same is true for some of the ostensible Mach quotes in the preface to the 9th German edition (1933) of the *Mechanik*.

¹⁵ A number of English speaking historians and philosophers of science have undertaken the attempt to show that Mach’s rejection of the theory of relativity was an internal consequence of his philosophy of science. This attempt not only failed because they tried to prove something non-existent, but also because it is based on grave misinterpretations and distortions of Mach’s ideas.

Joseph Petzoldt, Mach still spoke very positively about relativity theory. The *Optik* preface, however, is dated "July 1913".

It is Ludwig Mach's fatal personality that plays a decisive role in this forgery. Ludwig, as an intern, had only spent a few days as an operational assistant in the surgical ward at the university hospital in Vienna when he realized that he was not cut out for practical medicine.¹⁶ A three year period followed, during which he worked for the optical firm Zeiss in Jena. A patent for a light metal alloy (Magnalium), which can be traced back to one of his father's ideas, made him very soon a rather wealthy young man. A young man, however, who was criticized for being, among other things, a secrecy monger. He even went so far as to hide his marriage from his parents over a period of years. Ludwig Mach was a fairly good experimenter and instrument maker, although he had practically no knowledge of mathematics or theoretical physics. This lack was severely criticized by his father, but without having much effect. With the money he saved Ludwig built a house in Vaterstetten near München in 1912 and in spring 1913 his parents moved in. This house stood on an isolated piece of wooded land. The two Machs planned additional experiments together, particularly the completion of the father's unfinished projects. They especially wanted to write a book about primitive technologies, which appeared in 1915.¹⁷

The relationship between the two Machs is characterized by a strange dialectic of dependency and dominance. The elder Mach was dependent on his medically trained son in many ways since he could only move about the house or in the garden with a great deal of effort and could only conduct experiments with Ludwig's help. On the other hand, Ludwig's lack of knowledge in mathematics and physics together with his considerable scientific ambitions made him dependent on his father. Most importantly, Ludwig's identity began to reduce itself more and more to that of a son of Ernst Mach. Such a borrowed identity may have secured

¹⁶ Many years later, usually whenever he sought financial assistance, Ludwig maintained that he gave up his medical career *because of* his father's stroke. Now this, however, is demonstrably false, for the simple reason that Ludwig had given up his medical career almost a year before his father's stroke. With his own version he undoubtedly wanted to point out his in fact not insignificant sacrifice for his father as well as explain why he was as a doctor penniless at certain points. It is admittedly also clear that Ludwig, I am afraid one has to say it this way, lied a lot, but was not an especially good liar. For in order to help the father, after his stroke it would have been appropriate for Ludwig not to break off, but to recontinue his medical career.

¹⁷ E. Mach, *Kultur und Mechanik* (W. Spemann, Stuttgart, 1915) (Reprint Amsterdam 1969). Ludwig's co-authorship of this book is strangely enough never mentioned.

the esteem of the surrounding world, but it also made demands on Ludwig that he was in no way capable of dealing with. Ludwig's dependency on his father also had its negative side. As a result Ludwig began to dominate his physically severely disabled, but mentally completely alert father. For someone standing outside this relationship the presumptuousness of Ludwig's behavior toward his father is revealed most clearly in his decision not to tell his father about the outbreak of World War One in August 1914. In order to accomplish this aim, he told others not to mention the war in their letters and simply confiscated all the letters in which those uninformed about his censorship wrote about it. There were also other unauthorized activities. In 1915 (not in 1913 as the dating of the *Optik* preface suggests) and apparently against his father's expressed intentions, Ludwig Mach made secret arrangements with the publisher to begin printing the *Optik*. Printing was thus completely under way prior to Mach's death on February 19th, 1916.¹⁸ In October 1915, Ludwig wrote to Paul Carus, the director of the Open Court Publishing Company in Chicago, asking whether he would permit a dedication for the *Optik*.¹⁹ Carus seems to have forgotten to answer the inquiry. The dedication and the ostensible date (July 1913), moreover, are just as forged as the *Optik* preface and its ostensible date. The publication of the *Optik*, however, soon ran into problems. Particularly because Ludwig became involved in the production of airplanes in Berlin at the end of 1916.

The end of World War One meant for Ludwig the loss of his fortune, which he had invested for the most part in federal bonds. In the wake of these events Ludwig tried to reconcile the administration of his father's literary legacy with his desire to be admired for his own scientific achievements and the necessity of earning money for himself and those dependent on him. Ludwig's own scientific pretensions were connected in particular with the Michelson experiment. As he himself was the co-creator of a new type of interferometer, he felt that he was capable of reproducing this experiment. In his correspondence, however, he also indicates that he was not in a position to theoretically interpret a positive

¹⁸ The manuscript's first 8 (out of 14) chapters were at the publishers by November, 1915; by March, 1916 the publisher had already received back a large part of the proofs. Ludwig's statement in one of the footnotes in the *Optik* preface that "printing began in Summer, 1916", i.e. after Mach's death, is not true. Copies of the correspondence with the publisher, along with other documents, are, when not mentioned otherwise, in the Philosophical Archive at the University of Konstanz (Federal Republic of Germany).

¹⁹ The original is in the possession of the Open Court Archive, Carbondale, Illinois.

result. It seems as if Ludwig had already thought of reproducing the Michelson experiment in May 1914. At that time, however, in *agreement* with the positive evaluation of relativity theory his father had communicated to Joseph Petzoldt. Only toward the end of the war did Ludwig seem to give in again to his pretensions. He did this precisely at the moment when he thought Friedrich Adler would take over the theoretical role that had previously been reserved for his father. As far as Ludwig was concerned it was not important that the perspective on this work would necessarily have to shift from agreement to rejection. For the mentor he had in view (i.e. Adler) had emerged meanwhile as a critic of the special theory of relativity. Ludwig held no convictions in theoretical physics because he had virtually no knowledge of this field. All he wanted to do, regardless of the situation, was to put his interferometer into operation as a means of increasing his fame and relieving his bitter and, as we shall see, chronic financial misery by earning some money. Particularly in order to relieve his financial misery, Ludwig indicated that he still had important experiments to conduct for his father before he could take a stand on relativity theory "from his point of view". He never reveals what stand his father actually took on relativity theory. Until shortly before writing the *Optik* preface in the early part of 1921 Ludwig continued to state that he possibly would have something to say from *his father's point of view*, but he never announced that he would come about with something his father *actually* said. On the contrary, Ludwig as his father's designated heir, caretaker, and interpreter first had to conduct his experiments. *Only then*, he contended, could he make a statement, not his father's but from his father's point of view.

After being released from prison²⁰, Adler turned again to politics and was lost to Ludwig as theoretical mentor. He was replaced by the anti-relativity philosopher Hugo Dingler (1881–1954). In the meantime Ludwig had lost some of his interest in reproducing the Michelson experiment. The British solar eclipse expedition in 1919, however, motivated him to begin experimenting again. This expedition had confirmed what the general theory of relativity had predicted all along: that light bends when it passes through the sun's gravitational field. This expedition gave Ludwig the idea of using his interferometer to try to measure the bending of light by such small masses as tree trunks, the sides of houses

²⁰ Adler had been imprisoned since 1916 in connection with the murder of the Austrian prime minister. He was released in 1918 at the end of the war.

or similar objects. A very foolish idea in view of the minimal effects produced. Under these circumstances Ludwig refrained from saying anything to Dingler about relativity theory, be it positive or negative. In fact, he was so persistent that in the early part of 1920 Dingler imagined himself to have fallen into disgrace with Ludwig Mach because of his own, openly proclaimed *anti*-relativism. By this time, though, Ludwig had already decided against relativity theory as a result of a difficult situation he had gotten into in connection with the publication of the 8th edition of the *Mechanics* in 1921. According to the last will of Ernst Mach, the publishers had entrusted Joseph Petzoldt (and not Ludwig Mach) with the new edition of the *Mechanics*. Petzoldt was planning on adding a pro-relativity afterword to the 8th edition as a way of celebrating Mach's pioneering role. At first Ludwig had nothing against this decision, that is, in 1919, before he had made the decision to work together with the anti-relativist Dingler. Only when he realized that in following his own urge to be an outsider he would have to join ranks with Dingler (whose theoretical help he expected) and the anti-relativists, did he see the necessity of protesting Petzoldt's afterword; and only in the course of the resulting interpersonal and legal conflicts did Ludwig begin to distance himself step by step from what he had been saying up to that point — that he could at most say something about relativity theory from his father's point of view and then only after his own experiments.

The last step in this retreat from the truth was reached with the foreword to the *Optics*. Here Ludwig had created the impression that his father had rejected the theory of relativity and announced that the arguments for this rejection would be supplied in a second volume of the *Optics*. After World War Two Ludwig, then 80 years old, confessed to Dingler that the second volume of the *Optics* was a phantom. A phantom whose promised realization was supposed to reveal Ludwig as Mach's intellectual heir and keep alive the interest among the Mach and anti-relativity fans. A phantom, with whose help Ludwig succeeded again and again in obtaining not insignificant sums of money that he used mostly to meet his daily needs. Over decades Ludwig lived on the border of bodily, psychological and financial ruin. As early as the 1920s, he seems to have been dependent on cocaine. In a continual state of desperation he was thankful for whatever help his father's friends gave him. Admittedly, none of these friends and helpers had ever read anything of Ludwig's on relativity theory, although they were very interested in this. This work never existed. As a means of compensating as it were for a lack of ability, Ludwig gave himself a doctorate in philosophy in 1935 in addition to his

honestly acquired medical degree. Strangely enough, the self-named “double doctor” seems never to have encountered any question in his immediate vicinity about how he acquired his second doctorate.

In short, what we are dealing with in Mach’s apparent rejection of the theory of relativity is a soap opera; in many respects this soap opera has been continuing right up to the present day. A harried and failed man led the people he knew as well as researchers around by the nose. Some of these researchers were highly pleased by this because they felt they had found in Mach’s ostensible rejection of the theory of relativity an easy argument against Mach’s philosophy of science. In the future one will have to look for better ones.

4. Mach and the atoms

The situation is quite different with respect to the existence of atoms. It is an incontestable fact that Mach refused to believe in the existence of atoms. This must be distinguished from his attitude toward scientific theories that make use of the theory concerning the atomic constitution of matter. One can leave open the question of the existence of atoms and still work theoretically with the hypothesis that atoms exist. The young Ernst Mach did this until the mid-1860s. He considered the theory of atoms compared with the dynamic continuum theory “as the only acceptable one in physics” because of its greater theoretical efficiency and simplicity (MACH 1863: p. 13). For the same reason, that is, with respect to the principle of economy, he later held atomic theories and constitution theories on the whole to be not very practical. He did not, however, reject them right from the start.²¹ For “every idea which can help, and does actually help, is admissible as a means of research” (MACH 1900: *Wärmelehre* 431 fn., *Theory of Heat* 445 fn. 9). The actual development of science decides whether this is the case or not. This is why Mach can make the following statement without a trace of polemic: “The atomistic philosophy has recently gained ground again owing to the advances made in stereochemistry” (MACH 1900: *Wärmelehre* 430, *Theory of Heat* 390).

Einstein may have contributed in an important way to the fact that

²¹ In particular Mach did not shift from a theory of atoms to a continuum theory: “One would also misunderstand me if one ascribed to me a preference for the assumption that space is continuous” (MACH 1900: *Wärmelehre* 431 fn., *Theory of Heat* 445, fn. 9).

significant changes took place in Mach's estimation of the theoretical efficiency or, as Mach would have said: economy of atomic theory in the following years. For in a famous article on Brownian motion likewise published in 1905, Einstein had demonstrated the fruitfulness of the atomic hypothesis with respect to a puzzling phenomenon and at the same time provided the possibility for an indirect proof of the existence of atoms. The work done by Jean-Baptiste Perrin supplied the necessary proof within a short period of time (EINSTEIN 1922: p. 63). During his above-mentioned visit to Mach in September, 1910, Einstein undoubtedly had a chance to speak about the theory of atoms.²² On this occasion Mach confirmed once again his standard view that he would accept the theory of atoms at the moment in which it proves to be "more economical" than "phenomenological" theories. By phenomenological theories he meant such theories that limit themselves to representing empirical facts without any speculative hypothesis about the constitution of matter. The phenomenological theory of thermodynamics is the standard example for such a theory. When Einstein left Mach he had formed the impression that Mach had clearly expressed himself more positively about the theory of atoms in conversation than he had in his writings.

Now for an instrumentalist like Mach accepting a theory in no way implies the acceptance of the corresponding ontology, which means here assuming the existence of atoms. On the other hand, of course, "real" entities also require a corresponding theory. It seems that Mach saw himself forced to believe in the existence of atoms before he was able to convince himself of the economical value of the theory of atoms. This brings us to an episode in Mach's life that has puzzled quite a few scholars.

The Viennese physicist Stefan Meyer reported in 1950 that he once demonstrated to Mach how a spinthariscopes works and that Mach, seeing the visible traces on the screen caused by α -particles, called out: "Now I believe in the existence of atoms" (MEYER 1950: p. 3). Until quite recently the prevailing view had been that the episode originally took place in 1903 (cf. BLACKMORE 1972: pp. 319ff.). New evidence now suggests that the episode should be dated in connection with the founding of the Viennese Radium Institute at the end of 1910 (BLACKMORE 1988). This date is correct for the following two reasons: (1) Mach called the atomic theory uneconomical and denied the existence of atoms up to 1910

²² For details see WOLTERS (1987: Section 9).

but not in his new writings after 1910.²³ (2) It is almost inconceivable that a thinker of Mach's caliber would have behaved so inconsistently without having a very good reason. The only thing in favor of the 1903 dating, by the way, is an easily recognizable misreading of Meyer's report.²⁴ On the other hand, it can easily be shown that the spintharoscope episode could not have occurred before 1909. For only at this time could it be demonstrated that each of the traces corresponds to an α -ray falling on the spintharoscope screen.

Another argument for a fundamental shift in Mach's views on the theory of atoms can be found in his correspondence from 1914. Writing to an Austrian chemist in June 1914, Mach indicated that he had re-evaluated his first (1872) extended work in history and philosophy of science, *Die Geschichte und die Wurzel des Satzes von der Erhaltung der Arbeit* (MACH 1909). Mach viewed this work, which had a special significance for him since it was the foundation of his more mature work in the philosophy of science, "as not corresponding to reality, as *outdated* and *eccentric*".²⁵ We can deduce with a high degree of plausibility from the context of the letter that Mach alludes here to progress in atomic theories and it is these which correspond to reality. In other words, atoms really exist. That Mach repudiated his earlier conceptions as "eccentric" is very likely connected with one of his later views. Namely that he was presumptuous enough to make his principle of reality (see Section 2) so concrete that it excluded *a priori* as it were thoroughly possible experiences. The reason for this error lay in a dogmatic notion of the concept of observation. Mach did not take into consideration that indirect observations could also fulfil his reality principle. He did not realize that even the meaning of the concept of observation could change in the process of scientific progress. Mach's failure in this respect, however, also reveals his greatness as a human being and scientist. At 76 he still had the inner freedom to admit a mistake and to reject as erroneous, and even "eccentric" a view he had held during his best years.

5. Final remarks

Mach's attitude to his own mistakes not only reveals his stature as a human being, it also displays an essential trait of his philosophy of

²³ He probably retained his early position in the 7th German edition of the *Mechanik* (1912) for reasons of historical faithfulness.

²⁴ See also for the following WOLTERS (1988).

²⁵ Letter to Rudolf Orthner. Original in the Upper Austrian State Museum in Linz.

science, namely its *historical*, or what one could perhaps also call its *dialectical* character (cf. WOLTERS 1986, 1989).

According to Mach we learn from history *first* that things are constantly changing and that the seemingly absolutely certain results of science cannot stop this process: "In fact, if one learned from history nothing more than the changeability of views it would already be invaluable. The Heraclitean phrase is true for science more than for any other thing: 'One cannot enter the same river twice!' The attempt to retain the beautiful moment through text books has always been in vain. In time one should become accustomed to the fact that science is unfinished, changeable" (MACH 1909: p. 3). Mach's point is that science *as such* is "unfinished, changeable". There is no indication that he would have exempted the science of his day from his view of science as inevitably tentative and incomplete. However, history teaches us more than that everything at all times is in flux. This is Mach's *second* insight. We also learn from history to understand better what we think we know as "true" (at a particular historical point in time): "A view, of which the origin and development lie bare before us, ranks in familiarity with one that we have personally and consciously acquired and of whose growth we possess a very distinct memory" (MACH 1900: *Wärmelehre* 1, *Theory of Heat* 5). The continuation of this quote shows, finally, why Mach is so interested in history. For history supplies, and this is Mach's *third* insight, the possibility of understanding the *progress of science*, and at the same time provides perspectives on what an individual himself has to do to bring about scientific progress. For, as Mach says in the continuation of the passage just quoted, a view whose origin and development we know "is never invested with that immobility and authority which those ideas possess that are imparted to us ready formed. We change our personally acquired views far more easily." The cardinal point of Mach's historical conception of science is thus to use our understanding of the past in order to gain a perspective on how we can shape the future.

In the process Mach did not lose sight of the dialectic between scientific and technological progress on the one hand and social progress on the other. He realized this long before the qualitative shift in both processes actually occurred. As far as scientific and technological progress are concerned, we should not forget that the earth's resources are not inexhaustible: "The humming of streetcars, the whirring of factory wheels, the glow of the electric light, all these we no longer behold with unalloyed pleasure if we consider the requisite amount of coal burned every hour. We are fast approaching the time when these hoards, built up as it were when the earth was young, will in the old age of the earth have

become nearly exhausted. What then? Shall we sink back into barbarism? Or will mankind by then have acquired the wisdom of age and learnt to keep house?" (MACH 1906: *Erkenntnis und Irrtum* 80, *Knowledge and Error* 58).

With respect to the dominant economical structures, Mach believed — and to a certain extent in opposition to historical materialism — that it would be the progress achieved in *intellectual* culture and not changes in the economical structures that would lead the working class to “recognize the true state of affairs and confront the ruling sections (of society) with the demand for a use of the common stock of property that is more just and more appropriate.”²⁶ Once again, Mach has not failed to recognize this special dialectic, this time of *social* progress. More than a decade before the founding of the first socialist state Mach pointed out that in the preferred socialist state of the future the governmental “organization should be confined to what is most important and essential, and for the rest the freedom of the individual should be preserved. In the contrary case slavery might well become more general and oppressive in a social democratic (i.e. socialist) state than in a monarchy or oligarchy” (MACH 1906: *Erkenntnis und Irrtum* 81 fn., *Knowledge and Error* 63 fn. 21). In precisely this way, one of the Soviet contributions to the previous International Congress of Logic, Methodology, and Philosophy of Science called “antifanaticism and antiauthoritarianism” a prerequisite for and a sign of the “valid ethos of valid science”.²⁷ I believe that such an orientation toward freedom and tolerance will strengthen the contribution to modern western philosophy that philosophy in socialist countries — and not just these — will make in the future. We all need this contribution of a great humanistic tradition with its enormous intellectual resources. We need it to solve the problems that Mach prophetically addressed and which concern us all. Problems, on which the survival of all of us depends.

Despite all of his sceptical insight into the dialectic of technological and social progress, Mach was in one respect an *optimist*. He believed resolutely in the possibility of peace between peoples. As we all know, Mach conducted his pioneering experiments on the velocity of sound by

²⁶ MACH 1906: *Erkenntnis und Irrtum* 80f. The English rendering of the German word “billig” as “cheap” (*Knowledge and Error* 58) is wrong in the present context. The correct meaning is “just”.

²⁷ FROLOW (1983: p. 227). The fact, however, that a large number of authors in this volume feel they have to demonstrate their orthodoxy by quoting the works of Marx, Engels and Lenin, not to mention the writings of that eminent philosopher Leonid I. Breshnew, shows that there is still a lot to do in terms of “anti-authoritarianism”.

using projectiles shot from weapons. In 1887, at the end of a paper he gave on these experiments, he talked about the possibility of peace and the uses to which the devices he used in his experiments could be put: "at times one completely forgets what horrible aims these contrivances serve." With bitter irony he then condemns the view²⁸ that "eternal peace is a dream and not even a beautiful dream": "We [...] can understand the soldier's fear of becoming ineffective through too long a peace. But it really takes a strong belief in our inability to overcome the barbarism of the Middle Ages not to hope for or expect any basic improvement in international relations." Mach continues with an optimism that in Europe would soon be falsified twice in the most horrible way: "Next to the questions that separate the peoples there emerge one after the other, ever clearer and stronger, the great, common goals that will abundantly claim all of man's powers in the future."

At about the same time a great Russian author radiated the same optimism, although from religious motives. In his "Speech against the War" Leo Tolstoy writes: "And this is why, though our powers may appear so insignificant in comparison with the powers of our opponent, our victory is just as certain as the victory of the rising sun's light over the darkness of night" (TOLSTOI 1936: 119f.).²⁹ We must be especially thankful for the efforts coming from this country which contribute to making Mach's optimistic desire for peace, like that of his great Russian contemporary, somewhat less utopian in these days.

(Translation Steven Gillies)

References

- BLACKMORE, J.T., 1972, *Ernst Mach: His Work, Life and Influence* (University of California Press, Berkeley, Los Angeles, London).
- BLACKMORE, J.T., 1988, *Mach über Atome und Relativität – neueste Forschungsergebnisse* in: R. Haller and F. Stadler, eds., *Ernst Mach – Werk und Wirkung* (Hölder, Pichler, Tempsky, Wien), pp. 463–483.
- BUTTS, R.E., 1973, *Whewell's Logic of Induction*, in: R.N. Giere and R.S. Westfall, eds., *Foundations of Scientific Method: The 19th Century* (Bloomington Ind.), pp. 53–85.
- COHEN, I.B., 1955, *An Interview with Einstein*, *Sci. Am.* 193, pp. 69–73.
- EINSTEIN, A., 1922, *Untersuchungen über die Theorie der 'Brownschen Bewegung'*, R. Furth, ed. (Akademische Verlagsgesellschaft, Leipzig).
- FROLOW, I., 1983, *Soziologie und Ethik der Erkenntnis, des Lebens und des Menschen*, in:

²⁸ This view is attributed to the unfortunately unnamed "most important warrior and silent man of our time" (Bismarck?) (MACH 1925: p. 380ff.).

²⁹ I am grateful to PD Dr. Sebastian Kempgen (University of Konstanz) for helping me with the Russian original of the Tolstoy text.

- Institut der Philosophie der Akademie der Wissenschaften der UdSSR ed., Logik, Methodologie und Philosophie der Wissenschaft (Moscow), pp. 212–227.
- HERNECK, F., 1986, *Einstein und sein Weltbild. Aufsätze und Vorträge*, 3rd edn. (Der Morgen, Berlin (GDR)).
- KUZNETSOV, B.G., 1977, *Einstein. Leben–Tod–Unsterblichkeit* (Birkhäuser, Basel, Stuttgart).
- MACH, E., 1863, *Compendium der Physik für Mediciner* (Wilhelm Braumüller, Wien).
- MACH, E., 1900, *Die Principien der Wärmelehre. Historisch-kritisch entwickelt*, 2nd edn. (J.A. Barth, Leipzig); 1986, Engl. *Principles of the Theory of Heat. Historically and Critically Elucidated*, introd. M.J. Klein, ed. B. McGuinness (Reidel, Dordrecht).
- MACH, E., 1906, *Erkenntnis und Irrtum. Skizzen zu einer Psychologie der Forschung*, 2nd edn. (J.A. Barth, Leipzig); 1976, Engl. *Knowledge and Error. Sketches on the Psychology of Enquiry*, introd. E.N. Hiebert (Reidel, Dordrecht, Boston).
- MACH, E., 1909, *Die Geschichte und die Wurzel des Satzes von der Erhaltung der Arbeit*. Vortrag, gehalten in der K. Böhmisches Gesellschaft der Wissenschaften am 15. Nov. 1871, 2nd edn. (J.A. Barth, Leipzig).
- MACH, E., 1921, *Die Prinzipien der physikalischen Optik. Historisch und erkenntnispsychologisch entwickelt* (J.A. Barth, Leipzig); 1926, Engl. *The Principles of Physical Optics. An Historical and Philosophical Treatment* (London).
- MACH, E., 1923, *Die Analyse der Empfindungen und das Verhältnis des Physischen zum Psychischen*, 9th edn. (G. Fischer, Jena); Reprint, pref. G. Wolters (Wissenschaftliche Buchgesellschaft, Darmstadt, 1985); 1959, Engl. *The Analysis of Sensations and the Relation of the Physical and the Psychical*, introd. T.S. Szasz (Dover, New York).
- MACH, E., 1933, *Die Mechanik in ihrer Entwicklung historisch-kritisch dargestellt*, 9th edn. (Brockhaus, Leipzig); 1974, Engl. *The Science of Mechanics: A Critical and Historical Account of its Development*, 6th edn. (Open Court, La Salle IL).
- MEYER, S., 1950, *Die Vorgeschichte der Gründung und das erste Jahrzehnt des Institutes für Radiumforschung*, in: Festschrift des Institutes für Radiumforschung anlässlich seines 40jährigen Bestandes (1910–50), in: Sitzungsberichte der österreichischen Akademie der Wissenschaften, math.-naturwiss. Kl., Vol. 159, Heft 1/2 (Wien), pp. 1–26.
- MITTELSTRASS, J., 1972, *Das praktische Fundament der Wissenschaft und die Aufgabe der Philosophie* (Universitätsverlag, Konstanz).
- PLANCK, M., 1965, *Die Einheit des physikalischen Weltbildes* (Vortrag, gehalten am 9. Dezember 1908 in der naturwissenschaftlichen Fakultät des Studentenkörpers der Universität Leiden, in: M. Planck, Vorträge und Erinnerungen, 7th edn. (Wissenschaftliche Buchgesellschaft, Darmstadt) pp. 28–51.
- TOLSTOI, L., 1936, *Oeuvres complètes*, ser. I, Vol. 38, V. Tchertkoff ed. (Moscow); 1968, Germ. *L.N. Tolstoj, Rede gegen den Krieg. Politische Flugschriften*, P. Urban, ed. (Insel, Frankfurt), pp. 163–170.
- WITTICH, D., 1986, *Zur Entstehungs- und Rezeptionsgeschichte von W.I. Lenins Werk "Materialismus und Empirio-kritizismus"*, Sitzungsberichte der Sächsischen Akademie der Wissenschaften zu Leipzig 127, Heft 2.
- WOLTERS, G., 1986, *Topik der Forschung. Zur wissenschaftstheoretischen Funktion der Heuristik bei Ernst Mach*, in: C. Burrichter, R. Inhetveen and R. Kötter, eds., *Technische Rationalität und rationale Heuristik* (Schöningh, Paderborn), pp. 123–154.
- WOLTERS, G., 1987, *Mach I, Mach II, Einstein und die Relativitätstheorie. Eine Fälschung und ihre Folgen* (de Gruyter, Berlin, New York).
- WOLTERS, G., 1988, *Atome und Relativität—was meinte Mach?*, in: R. Haller and F. Stadler, eds., *Ernst Mach – Werk und Wirkung* (Hölder, Pichler, Tempsky, Wien), pp. 484–507.
- WOLTERS, G., 1989, *Visionary Positivism: Ernst Mach's Philosophy of Science* (Scientia, Bologna).

CONTRIBUTED PAPERS

Section 1. Foundations of Mathematical Reasoning

- M.A. ABASHIDZE, Ordinally Complete Normal Extensions of the Logic of Provability
S.N. ARTEMOV, On Logical Axiomatization of Probability
S.C. BAILIN, Analysis of Finitism and the Justification of Set Theory
L.D. BEKLEMISHEV, On the Cut-Elimination and Craig Interpolation Property for Probability Logics.
S.M. BHAVE, The Fallibilist Methodology of Mathematics
B.R. BORICIC, Some Intermediate Logics as Natural Deduction Systems
S. BUZASI, A.G. DRAGALIN, A Machine-Oriented Version of Analytical Tableaux Method
J. COUTURE, Proof Theory and Meaning
G.K. DARJANIA, Complexity of Deduction and Complexity of Countermodels for Grzegorzczyc's and Kuznetsov's Modal Systems
J.C. DE OLIVEIRA, Thirdness as a Common Limitation in Contemporary Mathematics
J. DILLER, Syntax of E-logic
A.G. DRAGALIN, Completeness and Strong Normalization in Higher-Order Logics. A Constructive Proof
P. FILIPEC, Constructive Versus Classical Mathematics
Y. GAUTHIER, The Internal Logic of Local Structures: A Representation Theorem For Local Negation and Complementation
A. GIUCULESCU, The Mathematicity of Scientific Theories
V. GOMEZ PIN, Reasoning in the Category of Quantity After the Non-Standard Analysis
S.V. GORYACHEV, On Relative Interpretability of Some Extensions of Peano Arithmetic
A. GRIEDER, On the Genetic Approach in the Philosophy of Geometry
L. GUMANSKI, On the Diagonal Method
W. HEITSCH, A New Algorithm for the Design of Optimum Binary Search Strategies
G. JÄGER, Non-Monotonic Reasoning By Axiomatic Extensions (invited speaker)
G.K. JAPARIDZE, Generalized Provability Principles and Modal Logic
M.I. KANOVICH, A Strong Independence of Invariant Sentences
C.F. KIELKOPF, The Mathematical A Priori After Kitcher's Critique
SHIH-CHAO LIU, A Finitary Proof of the Consistency of N and PA
V.A. LUBETSKY, Interpretation of Heyting Algebras Morphisms in Heyting-Valued Universum
L.L. MAKSIMOVA, Craig's Interpolation Property in Propositional Modal Logics
D. MARGHIDANU, Le Comportement des Opérateurs sur des Formules Propositionnelles Complexes
E. MAULA, E. KASANEN, Fermat's Heuristics
G.E. MINTS, Cutfree Formalizations and Resolution Method for Propositional Modal Logics
J. MOSTERIN, How Set Theory Impinges on Logic
N.M. NAGORNY, On Presentation of Elementary Semiotics
I.S. NEGRU, On Some Algebraic Applications of the Totalities of Logics
N.N. NEPEIVODA, Constructive Logics

- E. Yu. NOGINA, Some Algorithmic Questions in Lattice of Extensions of Finitary Logic of Provability
- V. PAMBUCCIAN, Euclidean Geometry is Axiomatizable by Sentences About up to Five Points
- J. PARIS, A. VENCOVSKA, Inexact and Inductive Reasoning (invited speaker)
- H. PFEIFFER, A Generalized Version of Kruskal's Theorem and the Limits of Its Provability
- R.A. PLIUSKEVICIUS, On Applications of Skolemization for the Constructive Proof of Cut Elimination From Some Nonpredictive Calculi of Temporal Logics
- A. PRELLER, N. LAFAYE DE MICHAUX, Equality of Objects in Categories with Structure
- RUY J.G.B. DE QUEIROS, Note on Frege's Notions of Definition and the Relationship Proof Theory vs. Recursion Theory
- V.M. RUSALOV, Foundations of a Special Theory of Human Individuality
- V.V. RYBAKOV, Problems of Admissibility and Substitution, Logical Equations and Restricted Theories of Free Algebras (invited speaker)
- P. SCOWCROFT, Recent Work on Constructive Real Algebra
- A.L. SEMENOV, The Lattice of Logically closed Classes of Relation
- N.A. SHANIN, On Finitary Development of Mathematical Analysis on the Base of Euler's Notion of Function
- S.V. SOLOVYEV, On Decreasing the Formula's Depth in Proof Theory and Category Theory
- A. TAUTS, Higher Order Formular in Infinitary Intuitionistic Logic
- J.P. VAN BENDEGEM, A Finitist Treatment of the Real Numbers
- V.A. VARDANYAN, On Provability Resembling Computability
- S.N. VASSILYEV, V.I. MARTYANOV, V.M. MATROSOV, Computer Methods of Theorem Synthesis and Proving
- A.A. VORONKOV, Constructive Logic: A Semantic Approach
- E.W. WETTE, Kolmogorov Bounds to Consistent Information, and Ether-Geometry
- P. WOJTYLAK, Independent Axiomatizability in Intuitionistic Logic
- XIE HONG XIN, On the Theory of two Levels of Mathematical Foundation
- A.D. YASIN, On a Semantic Approach to the Notion of the Intuitionistic Logical Connective
- ZHA YOU-LIANG, Chunk and Production Pattern

* * *

Yu. A. MITROPOLSKII, M.I. KRATKO, V.N. SHEVELO, Printsip posledestviya Pi-kara-Volterra i yego rol v sovremennom matematicheskom yestestvoznanii

Section 2. Model Theory

- M. BALAIS, A Simplified Method of Evaluation in the First Order Predicate Logic
- S.A. BASARAB, Relative Elimination of Quantifiers for Henselian Fields
- A. BAUDISCH, Classification and Interpretation
- O.V. BELEGRADEK, On Groups of Finite Morley Rank
- V.Y. BELYAEV, On Algebraic Closure and Embedding Theorems of Semi-groups
- A. ENAYAT, Trees and Power-Like Models of Set Theory
- L. ESAKIA, A Classification of the Elements of the Closure Algebras — Hausdorff's Residues
- R. Sh. GRIGOLIA, Free Products of Closure Algebras

- GUO JIN BIN, What is the Meaning of Mathematical Model
 R. GUREVICH, On Symbolic Manipulation of Exponential Expressions
 V.I. HOMIC, On Some Properties of Generalizations of Pseudo-Boolean Algebras
 L. KIUCHUKOV, System and Theory: The Ideals of the Constructive Style of Cognition
 A.I. KOKORIN, Mathematical Model of Transition From Polytheism to Monotheism
 M. KRYNICKI, J. VÄÄNÄNEN, Henkin Quantifiers and Related Languages
 M. MAKKAI, Strong Conceptual Completeness for First Order Logic
 V.J. MESKHI, Injectivity in the Variety of Heyting Algebras With Regular Involution
 A.S. MOROZOV, Permutations of Natural Numbers and Implicit Definability
 R. MURAWSKI, Definable Expansions of Models of Peano Arithmetic
 T.G. MUSTAFIN, On Some Properties of Stable Theories of Acts
 T.A. NURMAGAMBETOV, A Property of Models of Nonmultidimensional Theories
 E.A. PALYUTIN, Quasivarieties With Nonmaximum Spectrum
 A.G. PINUS, Skeletons of Congruence Distributive Varieties
 P. ROTHMALER, A Rank for Ordered Structures
 S.V. RYCHKOV, Application of the Axiomatic Set Theory to Group Theory
 K. TAGHVA, Preservation of Model Completeness under Direct Power in an Extended Language
 M.A. TAITSLIN, Dynamic Logics
 L. VAN DEN DRIES, The Logic of Weierstrass Preparation (invited speaker)
 A. VINCENZI, On the Abstract Model-Theoretic Neighbourhood of the Logics of Computers Languages
 A. WILKIE, J. PARIS, On the Existence of End Extensions of Models of Bounded Induction (invited speaker)
 H. WOLTER, On Ordered Exponential Fields
 B.I. ZILBER, Towards the Structural Stability Theory (invited speaker)

Section 3. Foundations of Computation and Recursion Theory

- V.M. ABRUSCHI, G. MASCARI, A Logic of Recursion
 M.M. ARSLANOV, Completeness in Arithmetic Hierarchy and Fixed Points
 N.V. BELYAKIN, L.N. POBEDIN, Dialogue Procedures in the Foundations of Mathematics
 V.K. BULITKO, On Setting Algorithmic Behaviour by Optimum's Principles
 M.S. BURGIN, The Notion of Algorithm and Turing-Church's Thesis
 V.E. CAZANESCU, Ch. STEFANESCU, A Calculus for Flowchart Schemes
 R. CUYKENDALL, Program Measures and Information Degrees
 M. DE ROUGEMONT, Programs, Recursions and Intentions
 O. DEMUTH, Some Properties of Sets Interesting From the Point of View of Constructive Mathematics
 S. FEFERMAN, Recursion-Theoretic Analogues Between Algebra and Logic (Formulations and Questions)
 P.N. GABROVSKY, On the Role of Selection Operators in General Recursion Theory
 F.W. GORGY, A.H. SAHYOUN, On the Impossibility of Transformation of All True Formulas of Any One of the Languages of Markov's Hierarchy L into True Formulas of Any One of the Languages of Markov's Hierarchy
 Yu. GUREVICH, Logic and the Challenge of Computer Science
 O.G. HARLAMPOVICH, The Word Problem for Solvable Groups
 E. HERRMANN, Automorphisms of the Lattice of Recursively Enumerable Sets and Hypersimple Sets (invited speaker)

- A. ISHIMOTO, Axiomatic Rejection for the Propositional Fragment of Lesniewski's Ontology
- JIANMIN ZHEN, The "Big Structure Method" for the Study of the Technological Development
- C.G. JOCKUSCH, Degrees of Functions With no Fixed Points (invited speaker)
- M.I. KANOVICH, Lossless Calculi As a Tool to Reduce the Search for Analysis-and Synthesis Algorithms in the Knowledge Base Systems
- N.K. KOSSOVSKY, The Combinational Complexity and Logic-Arithmetical Equations
- J. KRAJICEK, Measures of Complexity of Proofs
- E.V. KRISHNAMURTHY, Non-Archimedean Valuation — Its Philosophy and Practical Utility for Rational Recursive Computation
- V.N. KRUPSKY, Along the Path of Constructing a Complexity Hierarchy for the Points of R_n
- A. KUCERA, An Alternative View On Priority Arguments
- B.A. KUSHNER, A Counterexample in the Theory of Constructive Functions
- H. LEVITZ and W. NICHOLS, A Recursive Universal Function For the Primitive Recursive Functions
- LIANG DADONG, The Relationship Between Human Thought and Artificial Intelligence
- V.L. MIKHEEV, Algorithmical Isol Structures
- D. MILLER, Hereditary Harrop Formulas and Logic Programming
- M.U. MOSHKOV, On the Problem of Minimization of Algorithms Complexity
- R. Sh. OMANADZE, Relation Between Recursively Enumerable Q - and T -Degrees
- P. PUDLAK, A Note on Bounded Arithmetic
- A.L. RASTSVETAEV, About Recognizability of Some Properties of Monadic Schemas of Programs With Commutative Functions
- J. RYSLINKOVA, Decidability of Monadic Theories and Rewriting Techniques
- V.Yu. SAZONOV, Gandy's Theorem in Kripke-Platek Theory With Classes and Type-Free Analogue of Ershov's Calculus of Σ -Expressions
- D. SEESE, Decidability of Monadic Second Order Theories
- V.L. SELIVANOV, Hierarchies and Index Sets
- L.V. SHABUNIN, Decidability of \exists -theories of Finitely Presented Groupoids
- N. SHANKAR, A Machine-Checked Proof of Gödel's Incompleteness Theorem
- E.Z. SKVORTSOVA, F-Reducibility and Arithmetical Hierarchy of Classes
- C.H. SMITH, Inductive Inference: A Mathematical Theory of Learning by Example
- D. SPREEN, Computable One-One Enumerations of Effective Domains
- A.P. STOLBOUSHKIN, Constructive Enrichments of Algebraic Structures
- D.I. SVIRIDENKO, On Problem of Semantic Programmes Design
- M.R. TETRUASHVILI, On the Complexity of Decision of the Validity Problem For the Elementary Sublanguage of the Quantifier Free Set Theory
- H. THIELE, Eine modelltheoretische Begründung analogen Schliessens
- P. URZYCZYN, A Remark on the Expressive Power of Polymorphism
- Ju.M. VAZENIN, Hierarchies and Critical Theories

* * *

- S.T. ISHMUKHAMETOV, O metode prioriteta s beskonечnymi narusheniyami na raznosyakh rekursivno perechislimykh mnozhestv

Section 4. Set Theory

- U. ABRAHAM, Free Sets for Commutative Families of Functions (invited speaker)
- J.A. AMOR, The Cantor's Continuum Problem as a Real Problem

- M. FOREMAN, A Dilworth Decomposition Theorem for λ -Suslin Quasi-orderings of \mathbb{R}
(invited speaker)
- J.E. BAUMGARTNER, Polarized Partition Relations and Almost-disjoint Functions
(invited speaker)
- V.Kh. KHAKHANIAN, Special Realizability in Set Theory
- A. OBERSCHELP, On Pairs and Tuples
- V.E. SHESTOPAL, A Set Theory with Understricted Comprehension
- A. SOCHOR, The Horizon in the Alternative Set Theory
- XIANG WUSHENG, Set and Truth
- ZHANG JINWEN, A Hierarchy of Axiom Systems for Set Theory

Section 5. General Logic

- A.M. ANISOV, L^H -Theories and the Generalized Completeness Problem
- O.M. ANSHAKOV, V.K. FINN, D.P. SKVORTSOV, On Logical Means of Formalization
of Plausible Inferences
- G.S. ASANIDZE, Über mögliche unabhängige Axiomensysteme für deduktiv-abgeschlossene Satzsysteme
- M. ASTROH, Logical Competence in the Context of Propositional Attitudes
- C.I. BAKHTIYAROV, Arithmetization of Assertoric and Modal Syllogistics
- LIN BANGJIN, The Motivation of Constructing Entailment Logic
- D. BATENS, Nomological Implication
- G. BEALER, On the Significance of Completeness Results in Intensional Logic
- M. BELZER, A Logic for Defeasible Normative Reasoning
- M.N. BEZHANISHVILI, Logical Omniscience Paradox Free Epistemic Propositional
Systems
- K. BIMBO, Some Remarks about Conditionals
- A.L. BLINOV, Two Ways of Constructing a Logic of Action
- V.A. BOCHAROV, V.N. STEBLETSOVA, Semantics of Potential and Actual Worlds and
Aristotle's Conception of Potential Being
- G. BOOLOS, On Notions of Provability in Provability Logic
- V.N. BRJUSHINKIN, A Logic for Urn Models
- W. BUSZKOWSKI, Erotetic Completeness
- P.I. BYSTROV, Logic with Temporal Parameter and Relevant Implication
- A.V. CHAGROV, Possibilities of the Classical Interpretations of Intuitionistic Logic
- L.A. CHAGROVA, On First Order Definability of Propositional Formulas
- B. CHENODOV, Dyadic Modal System of Order One CPD_1
- S.V. CHESNOKOV, The Effect of Semantic Freedom in the Relationships between
Denominations in the Logic of Natural Language
- J. CIRULIS, Logic of Indeterminacy
- N.C.A. DA COSTA, Logic and Pragmatic Truth (invited speaker)
- L. DE MORAES, On Jaskowski's Co-Discussive Logic
- J.K. DERDEN Jr., Fictional Discourse and Analytic Truth
- E. DÖLLING, Are There Objects Which do not Exist?
- W.V. DONIELA, The Principle of Identity as a Problem
- K. FINE, The Justification of Negation as Failure (invited speaker)
- V.K. FINN and T. GERGELY, Constructivity of Plausible Inferences and Its Role in a
Theory of Reasoning
- G.I. GALANTER, On Some Representations of Logic $S5$ and Its Extensions
- G. GARGOV and S. RADEV, Arguments and Strategies (Lukasiewicz Meets Polya)
- M. GELFOND, On the Notion of "Theoremhood" in Autoepistemic Logic

- I.A. GERASIMOVA, Reasoning on the Ground of Personal Knowledge
 Ju. G. GLADKIKH, Logic of Change: A Semantical Approach
 R. GOLDBLATT, First Order Spacetime Geometry (invited speaker)
 GONG QIRONG, Entailment Logic— a Development of Traditional Logic in Our Times
 I.N. GRIFTSOVA, The Role of Notions “Situation” and “Event” in Logical-Semantic Analysis of Sentences
 M. HAND, Inference and Strategy-Conversion
 K.G. HAVAS, Laws of Logic from the Point of View of Philosophy of Science
 HE YI-DE, The Modern Development of the Science of Logic
 W. HEITSCH, Logical Relations in Questionnaires
 HO YIN SEONG, A Dynamic Logic
 G. HOLMSTRÖM, Actions and Negations
 N. HUNT, The Fictions Legal Argument
 A.T. ISHMURATOV, Towards Logical Theory of Practical Reasoning
 A.A. IVIN, The “Explanation-Understanding” Logics
 Yu. V. IVLEV, New Semantics for Modal Logic
 A.A. JOHANSSON, Imperative Logic Based on a Concept of Admissibility for Imperatives
 N. KANAI, The Characterization of the Aristotelian Syllogistic by the Countable Models
 M. KANDULSKI, The Nonassociative Lambek Calculus of Syntactic Types
 E.F. KARAVAEV, Tense-Logic Semantics and Period Temporal Structures
 A.S. KARPENKO, Logic as Truth-Value
 GYULA KLIMA, The Inherence Theory of Predication: an Old Theory in a New Perspective
 M. KOBAYASHI, The Cut-Elimination Theorem for the Modal Propositional Logic S5
 V.N. KOSTYUK, Basic Epistemic Logics
 K.-H. KRAMPITZ, Definitionen des Existenzprädikates und leere Termini
 L. KREISER, Logical Hermeneutics
 A. KRON, Lattices Definable in Terms of Implication and Negation
 A.A. KRUSHINSKY, Indeterminacy of Translation and Semantics of the Chinese Language
 N.G. KURTONINA, E.G. SHULGINA, What Shall We Gain by Using Activity Approach in Logical Semantics?
 V.G. KUZNETSOV, Logical Reconstruction of Aristotle’s and Hegel’s Understanding of Motion by Logic of Change
 I.S. LADENKO, Logic of Mathematical Modelling
 E.E. LEDNIKOV, On One Variant of Epistemic Logic free of Logical Omniscience Paradoxes
 L.L. LEONENKO, The Logical Properties of Some Calculi of the Language of Ternary Description
 F. LEPAGE, A Characterization of A Priori Knowledge in Intensional Logic
 M. LEWIS, Defeasible Thinking, Defeasible Logic, and D-Prolog
 V. LIFSCHITZ, J. MCCARTHY, Non-Monotonic Reasoning and Causality
 M. LISTIKOVA, Consecution and Causality in Artificial Intelligence
 P. LUDLOW, Substitutional Quantification and the Semantics of the Natural Language
 A. MADARASZ, A Case for the Combination of Game Theoretical Semantics with the Conception of Semantic Value Gaps
 G. MALINOWSKI, M. MICHALCZYK, Q-Matrices and Q-Consequence Operations
 V.I. MARKIN, Semantics for de re Modalities
 F. MAXIMILIANO MARTINEZ, Nota sobre el Cuadrado Lógico y las Reglas de Derivación
 L.I. MCHEDLISHVILI, On Reconstruction of Aristotelian Syllogistic
 V.S. MESKOV, On the Completeness of Quantum Mechanics: Syntactical Analysis

- T. MIHALYDEAK, Truth-Functions in the Intensional Logic Accepting Semantic Value-Gaps
- E.A. MIKHEEVA, The Problem of a Basis for Finitely-Valued Logics
- M.S. MIRCHEVA, Knowledge and Action
- M. MOSTOWSKI, An Extension of the Logic with Branched Quantifiers
- A.A. MUCHNIK, N.M. YERMOLAEVA, Retro-Temporal Logic and Finite Automata
- A. Yu. MURAVITSKY, On Properties of Isomorphism between the Lattice of Provability Logic Extensions and the Lattice of Provability-Intuitionistic Logic Extensions
- M. OMYLA, Ontology of Situations in the Language of Non-Fregean Sentential Logic
- E. ORLOWSKA, Semantics of Knowledge Operators
- S.A. PAVLOV, Axiomatic Approach to the Theory of Denotation and Lesniewski's Calculus of Names
- L. PENA, The Calculus of Determinations
- J. PERZANOWSKI, Ontological Modalities
- L. POLOS, Semantic Representation of Mass Terms
- V.M. POPOV, On the Extensions of "Ockham's Syllogistics" System C2
- M.V. POPOVICH, Context and Vagueness of Sense
- S. RAHMAN, Remarks on the Notion of Dialogues
- M.F. RATSA, The Problems of Expressibility of Modal and Predicate Formulae
- M. SANCHEZ-MAZAS, A New Arithmetical Decision Method for Equivalent Deontic Systems
- G. SANDU, Dyadic Logic of Action
- O.F. SEREBRIANNIKOV, Elementary Proof of the Normal Form Theorem in Second and Higher Order Logic
- S.D. SHARMA, Quantum Mechanical Systems and Analogical Approaches
- V.B. SHENTMAN, On Some Two-Dimensional Modal Logics
- N.N. SHULGIN, The Formal Explication of the Concept of Movement
- E.A. SIDORENKO, Entailment as Necessary Relevant Implication
- E.A. SMIRNOV, Assertion and Predication. Combined Calculus of Propositions and Situations
- E.D. SMIRNOVA, Logical Entailment, Truth-Value Gaps and Gluted Evaluations
- O.A. SOLODUKHIN, Models with a Changing Universum
- W. STELZNER, Tautological Entailment, Negation and Assignment Changers
- R. STUHLMANN-LAEISZ, Completeness for Some Dyadic Modal Logics
- M. TABAKOV, The Philosophical Importance of Paraconsistent Logics
- D.D. TEVZADZE, On One of Church's Objections Against Russell's Theory of Descriptions
- M. TOKARZ, Pragmatic Matrices
- M.S. UNGUREANU, Phenomenology and Logical Semantics; Towards an Epistemology of Semantics
- I. URBAS, Paraconsistency and the C-Systems of Da Costa
- A.I. UYEMOV, Fundamental Features of the Language of the Ternary Description as a Logical Formalism in the Systems Analysis
- D. VAKARELOV, S4 and S5 together — S4 + 5
- M.K. VALIEV, On Deterministic PDL with Converse Operator and Constants
- V.L. VASYUKOV, Quantum Logic of Observables as Converse Semantical Problem
- R. VERGAUWEN, How to Montague Language to the World
- E.K. VOISHVILLO, Modal Syllogistics Founded on the Concepts of Relevant Logic: An Interpretation of Aristotle's Apodictic Syllogistics
- S. WEINSTEIN, Philosophical Aspects of the Theory of Proofs and Constructions (invited speaker)

- H. WESSEL, Dialtheismus: Mystik im logischen Gewande
 A. WISNIEWSKI, The Generating of Questions and Erotetic Inferences
 B. WOLNIEWICZ, Presuppositions for a Generalized Ontology of Situations
 J. WOODS, Is Dialectical Logic Necessary?
 K. WUTTICH, Gibt es eine deskriptive epistemische Logik?
 A. ZAKREVSKY, Implicative Regularities in Formal Models of Cognition
 S. ZELLWEGER, Notation, Relational, Iconicity, and Rethinking the Propositional Calculus
 ZHU SHUI-LIN, On Modern Logic
 R. ZUBER, Intensionality and Non-Bivalence

* * *

G.I. DOMB, Ob odnoi teorii istinnostno-funktsionalnykh provalov

Section 6. General Methodology of Science

- Zh.L. ABDILDIN, Dialectical Logics as the Methodology of Constructing Theoretical Knowledge
 M.N. ABDULLAEVA, Reflection Adequacy as a Problem of Truth Theory
 R. ABEL, The Anthropic Principle: A Mistaken Scientific Explanation
 A.O. ABRAMYAN, Mathematization as a Major Factor of the Scientific and Technical Progress Acceleration
 O.V. AFANASJEVA, Subject-Object Dialectics in Creative Process
 M.D. AKHUNDOV, S.L. ILLARIONOV, Problems of Development of Science and its Reflection in the Methodology of Scientific Knowledge
 G.N. ALEXEEV, Methodology of the Complex Investigation of the Techniques (Technological) Development and Integration of Social, Technical and Natural Sciences
 L.G. ANTIPENKO, On Discovery of Two Levels of Thinking-Logical and Object-Intentional
 A.N. ANTONOV, Scientific Tradition as a Form of Science Development
 M.G. ANTONOV and K.I. SHILIN, Ecological Method
 I.K. ANTONOVA, Thing Formed as a Key to Understanding the Thing in Making (K. Marx on the Reconstruction of the Scientific Theory Development)
 R.L. APINIS, Cognitive and Normative-Practical Thinking About an Object
 A.K. ASTAFYEV, Social Determination of Natural and Technical Sciences Interaction
 B.S. BAIGRIE, Scientific Rationality: Can We Reconcile Epistemological Respectability with Historical Fidelity?
 V.V. BAJAN, Determinism and the Problem of Reality in the Methodology of Science
 L.E. BALASHOV, Category Structure of Thinking and Scientific Cognition
 S. BALLIN, Discovery, Creativity and Methodological Rule-breaking
 A. BALTAS, A Strategy for Constituting an Althusserian Theory on the Structure and on the History of Physics
 A.N. BARANOV, V.M. SERGEEV, Argumentation as an Instrument of New Knowledge Foundation
 A. BARTELS, Physik, Biologie und der Begriff der "Natürlichen Art"
 L.B. BAZHENOV, Reduction as a Specifically Cognitive Operation
 V. BEKES, On the Missing Paradigm

- V.F. BERKOV, Problemology as Part of Science General Methodology
- I.V. BLAUBERG, E.M. MIRSKY, Organization of Knowledge as a Methodological Problem of Science
- E.D. BLYAHER, L.M. VOLYINSKAYA, Migration of Scientific Cognition and Methodology of Informative Transfers
- D.B. BOERSEMA, The Irrelevance of the Realism/Non-Realism Debate
- I.A. BOESSENKOOL, Neo-Cartesian Methodology (Linguistics, How Languages Act)
- J. BOKOSHOV, Conceptual Premises of Scientific Cognition and the Problem of Understanding
- V.I. BOLSHAKOV, On some Problems of the System Approach Development
- V.N. BORISOV, On the Explication of the Epistemic Categories of "Knowledge", "Belief", "Doubt", "Fallacy"
- B. BORSTNER, Why to Be a "Realist"?
- E.S. BOYKO, Methodological Premises of Research in Scientific Schools Activity
- M. BRADIE, Metaphors in Science
- C. BRINK, J. HEIDEMA, A Verisimilar Ordering of Theories Phased in a Propositional Language
- F. BRONCANO, Technological Possibilities as a Line of Demarcation Between Science and Technology
- M.V. BUCAROS, R.V. SAMOTEEVA, The Specific Character of the Artistic Information and Its Role in the Mastering of Reality
- S.P. BUDBAYEVA, Subjective Probability and Inductive Logic
- F. BUNCHAFT, Derivation by Continuity and by Limit in the Mathematical Manuscripts of Karl Marx (A Controversy)
- M.S. BURGİN, V.I. KUZNETSOV, Scientific Theory and Its Axiology
- J.M. BYCHOVSKAYA, Methodological Role of the Objects Conceptual Scheme in the Social Cognition
- V.D. CHARUSHNIKOV, The Empirical Conception of the Grounds of Mathematics and Its Narrowness
- CHENG JIANG FAN, The Model of Final Common Pathway in Life Science
- V. CHERNIK, Systematic Nature of Modern Scientific Revolution
- V.V. CHESHEV, Practical Technical Knowledge in Structure of Science
- A. CORDERO, Constructive Empiricism on Observation: Arbitrary or Incoherent?
- U. D'AMBROSIO, On the Generation and Transmission of Science
- R. DASKALOV, The Preference for Rationalistic Constructions in the Social Sciences—Thoughts on Weber and Schütz
- V.S. DAVIDOV, The Problem of Proving the Truth in Scientific Cognition
- R.T. DE GEORGE, The Moral Limitations of Scientific Research
- A.J. DEGUTIS, Principles of Subjectivity and Rationality in Action Explanations
- V.A. DEMICHEV, Object and Subject-Matter of Science
- W. DIEDERICH, Revolutions of the Structure of Science
- A.I. DIMITROV, On the Possibility of a Space Generative Grammar
- A. DOBRE, La Corrélacion Méthode – Programme dans l'Action Humaine Rationelle
- A.M. DOROZHKIN, Role of Delusion and Uncertainty in Science Methodology
- A. DRAGO, An Effective Definition of Incommensurability and its Theoretical Implications
- A.I. DRONOV, Exchange Factor as a System Invariant
- D.I. DUBROVSKY, Preproblem and ad-Problem Situations: to the Analysis of the Emerging of a New Problem in Scientific Knowledge
- J.H. DURNIN, Studies in Teaching for Creativity through Computers
- E.S. DYAKOV, The Explication of the Notions of a Theory as a Logical-Gnoseological Operation

- P.S. DYSHLEVYJ, *Special Science Model of the Studied Reality as a Form of Scientific Cognition (on the Material of Physics)*
- Y.L. EGOROV, *Systematism in Knowledge as a Methodological Problem*
- K.E. ERMATOV, *The Principle of Unity as a Scientific Methodology*
- E. ERWIN, H. SIEGEL, *Is Confirmation Differential?*
- E.A. EVSTIFEEVA, *Belief as an Object of Methodology of Science*
- A.F. FAIZULLAEV, *Algorithm in the Past, Present and Future*
- V. FILKORN, *Strong and Weak Methods (invited speaker)*
- M.A. FINOCCHIARO, *Science and Religion: Toward an Interactionist View*
- R. FJELLAND, *The Uses of Idealization in the Exact Sciences*
- H.J. FOLSE, *Realism and the Quantum Revolution*
- A. FRANKLIN, *Dirac Theory, Bayesianism, and the Duhem–Quine Problem*
- M.C. GALAVOTTI, *On the Causal Interpretation of Statistical Relationships*
- P. GALISON, *Traditions of Instruments and Traditions of Theories in High-Energy Physics*
- P. GÄRDENFORS, *Conceptual Spaces and Scientific Theories*
- R.V. GARKOVENKO, *On the Methodology of Research In the Development of Natural Scientific Knowledge*
- J.P. GARSIA BRIGOS, *Análisis Filosófico-Metodológico de la Interacción entre los niveles empírico y teórico en el proceso del conocimiento científico*
- L.D. GENERALOVA, *Methodological Aspects of Investigating Social Progress*
- I.G. GERASIMOV, *The Philosophic Conceptions of Scientific Investigation and the Criteria of a Scientist's Professional Qualification*
- S.S. GERDIKOV, *A Matrix Model of Scientific Explanation*
- A.D. GETMANOVA, *Negations in Classical and Non-Classical Logics*
- E. GIANNETTO, *Quantum Epistemology*
- S.N. GONSHOREK, *Computational Experiment as a Logico-Informational Process*
- A.A. GORELOV, *The Ideal of Science as a Holistic "Integrated-Varied" Harmonious System*
- V.G. GOROKHOV, *Theoretical Synthesis of Scientific-Engineering Knowledge*
- T.D. GORSKAYA, *On Classification of Dialectical Contradictions in Cognition*
- D.P. GORSKIY, *Real and Ideal Types*
- V.V. GRECHANY, *Value as a Methodological Concept*
- K.V. GRIGORJEV, N.S. GRIGORJEVA, *On the Specific Features of Social Systems*
- V.P. GRITSENKO, *Interconnection between the Discrete and Continual in the Scientific Cognition*
- M. GROPPOSILA, *Creativity—an Outcome of Interaction. Elaboration of a Pattern on Creativity*
- U.A. GUSEV, V.A. VASILYEV, *Heuristic Function of the Dialectical Logic in Scientific Research*
- J.M. GVISHIANI, *Impact of Global Modelling on Modern Methodology of Science (invited speaker)*
- K. HAHLEWEG, *Realism Versus Relativism: an Evolutionary Approach*
- M. HARRELL, *Contemporary Geometry*
- HE YU-DE, *Philosophical Thinking About the Problem of Information; Synopsis*
- HE ZUORONG, *Views on the Methodology of Social Sciences*
- R. HILPINEN, *Cognitive Progress and the Logic of Abduction*
- C.A. HOOKER, *Towards New Foundations for Evolutionary Epistemology*
- P.G. HORWICH, *Does Believing a Theory Take More Than Just Using It?*
- H. HÖRZ, *Development of Science as a Change of Types*
- P. HOYNINGEN-HUENE, *World Constitution as the Basis of Thomas Kuhn's Philosophy of Science*

- R.I.G. HUGHES, Structural Explanations in Physics
G.M.K. HUNT, Experimentation, Information and Hypothesis Choice
U.P. HÜTT, Reflexion as Methodological Device
PO-KEUNG IP, Knowledge Change, Corpus Revision and the Ambiguity of Scientific Testing
K.I. IVANOVA, On the Impact of Scientific Revolutions upon the Structure of Philosophical Knowledge
A.A. IVIN, Values: The Central Topic of Contemporary Epistemology
JING SHICHAO, Residing of Methodology in Knowledge Teaching
W.B. JONES, Toward Distinguishing Science and Technology
H.V. KARAMYSHEV, Materialist Dialectics as the Foundation of Scientific Methodology
V.N. KARPOVICH, The Relationship Between Mathematics and Natural Science from a Methodological Point of View
V.P. KAZARYAN, The Nature of Time and Natural Sciences
V.V. KELLE, Justification Problem in Systems Modelling
P.P. KIRSCHENMANN, Chance Structures and Other Accidentalities: Reductionist and Holistic Approaches
H.N. KNYAZEVA, Accidentality as a Form of Knowledge
A.A. KOKORIN, Methodological Reductionism in Science: the Main Causes and Tendencies
V.I. KOLODYAZHNYI, Methodological Function of Categories of Philosophy in the Logic of Active Cognition
S.N. KOSKOV, Complementarity of Semantic Conventions and Metaphors in the Language of Science
I.M. KREIN, Methodological Aspects of the Problem of Contact of Human "Intelligence" with Other "Intelligent" Systems
D.T. KRIVENKO, K.T. KRIVENKO, On the Logic of a Scientific Object Formation (An Activity Hierarchic Model)
A.A. KRUSHANOV, The Integration of Modern Scientific Knowledge: Possible Lines of Development
G. KUENG, The Classification of the Different Kinds of Reasoning
T.A.F. KUIPERS, A Decomposition Model for Explanation and Reduction
V.I. KUPTSOV, Peculiarities of the Formation of the Modern Picture of the World
O.D. KURAKINA, From a Synergy to the Synarchy: Synergetics as Initial Stage of a Universal Synthesis
P.P. LAKIS, Forecasting in Natural Sciences as a Progress of Cognition
R. LAYMON, Using Scott Domains to Explicate the Concept of Idealized Data
W. LAWNICZAK, Scientific Concept and Its Cultural Archetype
F.V. LAZARYEV, Interval Research Program of Science Knowledge Analysis
J. LEROUX, On the Structuralist View of Empirical Theories
G.D. LEVIN, Family Resemblances Method as a Form of Generalisation
LI XINGMIN, The Preservation of the Essential Tension Between Opposing Extremes: A Highly Effective Principle of the Epistemology and Methodology of Science
LI YUANGING, Model Theory and Its Application
V.S. LIBENSON, The Criteria of Efficiency in Science
E.K. LIEPIN, Categorical Justification of the Specific Features of Scientific Disciplines
LIN KEJI, On the Method of Raising Abstraction to Concreteness
N.A. LITCIS, Lobatschewsky Viewpoints on Logical and Methodological Foundations of Science
LIU JI, Engineering Methodology and Its Utilization in the Technical Fields
LIU XINQI, Analysis of Failures in Scientific Research

- LIU YAN YAN, On Scientific Content of Information
- LIU YUESHENG, XU YILING, Wu Xuemou and Pansystems analysis—a New Exploration of Interdisciplinary Investigation
- A.M. LIZ, Una Clasica Objeción Para una no Menos Clasica Propuesta: el Caso del Realismo Interno
- V.O. LOBOVIKOV, Methodologic Significance of Approximate Treatment of Evaluations and Preferences for Justification of Norms and Choice Acts
- D.G. LOSCHAKOV, Reductionism and Antireductionism in Personality Research: Dilemma or Synthesis?
- P. MAHER, Why Scientists Gather Evidence
- L.G. MALINOVSKY, Principles of Building Substantial Mathematics
- P.V. MALINOVSKY, Style of Scientific Thinking: Effect of the Polysystem
- M.A. MAMONOVA, A Transitional Type of Rationality of Thinking: the Problem of Methodological Analysis
- SI-WAI MAN, High Confirmability, Unification and Realism in Scientific Theories
- C.C. MARE, Symmetry and Asymmetry, Order and Disorder in Universe (the Asymmetry as Order Factor and Qualitative Diversification Factor of Being)
- E. MARQUIT, On the Relationship Between Linguistic Expression and Formal Dialectical Logic
- P.O. MARTIN, P.C.L. TANG, Incommensurability: Towards a Unified Theory
- R. DE A. MARTINS, Proposal of a Directive Non-Prohibitive Methodology
- A.A. MARTIROSYAN, On a Method of Scientific Discoveries Logical Study
- P. MATERNA, Attributes and Information
- V.I. METLOV, Antinomies, Foundations, Thing
- H. METZLER, Zur Struktur der Empirie der Methodologie
- N. MICHOVA, Inventing Meanings and Modifying Meanings-basic Processes that Underlie Scientific Knowledge
- V.V. MILASHEVICH, Is There a Generic Cognition Process?
- A.G. MIROIU, Modal Axiomatizations of Theories and Theoreticity
- C. MISAK, Truth and the End of Inquiry
- J. MISIEK, Against Relativism
- E.P. MOKHORYA, On the Contents and Heuristic Functions of the Principle of Unity of Conservation and Change in Physical Cognition
- A. MONIN, L. TSIMBAL, I. SHMELEV, P.P. SHIRSHOV, Oceanology and Philosophy (Oceanology a Complex Science)
- A.T. MOSKALENKO, Methodological Basis of Interdisciplinary Studies
- N.V. MOTROSHILOVA, On the Possibility of Applying Phenomenological Methods to Modern Science
- E.A. MOUTSOPOULOS, Method and Intentionality
- G. MUNEVAR, Relativism as the Foundation of Science
- V.V. NALIMOV, Probabilistic Calculus of Meanings
- E.M. NEKRASAS, Prescription and Description in Methodological Research
- T. NICKLES, Logics of Discovery
- A.L. NIKIFOROV, Semantic Approach to the Analysis of Understanding
- A. NISANBAYEV, Dialectics of Truth and Error in Formation and Grounds of Scientific Theory
- V.J.A. NOVAK, Z. MASINOVSKY, K. ZEMEK, The Law of Transition of Quantity to Quality and the Evolution of Organisms
- I.B. NOVIK, Philosophical and Methodological Foundations and Paradoxes of Systems Modelling
- V.M. NOVIKOV, L.I. CHERNYSHOVA, Reflexive Structures of Self-consciousness as Necessary Premise of Theoretical Reflection

- M. NOVOSJOLOV, Identity with a Measure of Transitivity
- V. NOVOTNY, V.J.A. NOVAK, The Systemic Approach as One of the Main Components of the Dialectical Materialists Conception
- A. NOWACZYK, The Concept of Concept and Its Applications to Non-Statement Theories
- V.Y. OBUKHOV, Two Trends in the Development of an Object and Their Theoretical Interpretation
- E.A. ORLOVA, The Bases for Cultural Change Study
- M.S. ORYNBEKOV, On the Scientific Knowledge as a Specific System
- B.J. PAKHOMOV, The Dialectical Principle of Development and Its Methodological Role in Contemporary Natural Sciences
- H. PARTHEY, Methodological Structure of Research Situations
- Z. PARUSNIKOVA, What is Objective Knowledge?
- I. PARVU, On the Epistemological Status of the Theories of Universal Grammar
- D. PEARCE, Conceptual Change and the Progress of Science (invited speaker)
- A.A. PECHENKIN, Conceptual Foundations of Scientific Knowledge: Classical and Modern Approaches
- J. PEJSA, Methodological Aspects of the Relationships of Information Sources, System Evaluation and Realization of Technological Innovations During the Process of CAD Introduction
- M. PERA, Dialectics and Scientific Truth
- V. Ya PERMINOV, On Fallibilistic Conception of Mathematical Proof
- J. PINKAVA, Modeling in Natural Sciences
- I.P. POPOV, The Concrete and General: Distinction, Opposition, Contradiction, Precision
- R.N. PROCTOR, Political Values and Scientific Priorities: Historical and Philosophical Perspective
- U.A. RADJABOV, On the Systems-Methodological Aspects of Science Dynamics
- A.I. RAKITOV, Philosophy of Science in the Age of the Computer Revolution
- V. RANTALA, Explaining Superseded Laws
- A.N. ROGERO, Logic, Rationality and Historical Integrity of Knowledge
- E.I. ROGULENKO, T.A. ROMANYCHEVA, Methodology of Scientific Cognition
- S.B. ROSENTHAL, Kuhn, Pierce and Scientific Method: a Revisit
- D. ROTHBART, Methodology, Meaning and Metaphor
- H. ROTT, On Relations Between Successive Theories
- M.A. ROZOV, What is Science Composed of?
- N.S. RYBAKOV, The World of Theory
- V.N. SAGATOVSKY, About Basic Stages of Scientific Cognition
- SHASHI SAHAY, Some Desiderata for a Third World Approach to the General Methodology of Science
- E.P. SEMENYUK, Science by the End of the 20th Century: Methodological Features
- A.S. SEREBRYAKOVA, The Scientific Thinking Style as a Means of Acquisition and Development of Knowledge
- M.L. SHAMES, Scientific Epistemology: A Psycholiterary Approach
- A.L. SHAMIS, Conceptions of Growth and Development of Knowledge in the Modern Methodology of Science
- D. SHAMSIDDINOV, The Laws of Nature and the Laws of Science
- L.N. SHAVERDOVA, The Method and Style Unity Problem in the Scientific Cognition
- A.P. SHEPTULIN, The Correlation of the Dialectical Method and Methods of Particular Sciences
- G.A. SHIRSHIN, The Search for Mediations Between Science and Practice as a Methodological Problem
- K. SHRADER-FRECHETTE, Scientific Method and the Objectivity of Epistemic Value Judgements (invited speaker)

- SHU WEIN-GUANG, *On the Thinking Way of Subsystem*
 V.A. SHUCKOV, *Has Methodological Scientific Analysis Any Chance to Liberate Itself From the Circle of Philosophy?*
 E.D. SHUKUROV, V.K. NICHANOV, *Mimicry of Understanding or "Othello Effect"*
 T.Y. SIDORINA, *The Structure of the "Network of Theories" and Peculiarities of Their Formation in Modern Macrophysics*
 H. SIEGEL, *Can Philosophy of Science be Naturalized?*
 V.V. SILVESTROV, *Dialectical Logics as Foundation of the Link Between the Humanities and Natural Sciences*
 N.D. SIMCO, *Science, Methodology, and 20th Century Irrationalism*
 M.S. SLUTSKY, *The Role of Philosophic-Methodological Component in the Scientific Cognition*
 L. SOFONEA, *The Meta-Concept of "One" in the Thought of Physics*
 L. SOFONEA, N. IONESCU-PALLAS, I. GOTTLIEB, T. TORO, *Paradoxes and Counter Examples in the Modal Logic of Physics (Basic Items Only)*
 A.S. STEFANOV, *Self-Referential Inconsistency and the Refutation of Theories*
 V.F. STEPANENKO, A.N. DEDENKOV, *The Problem of Unity of Scientific Knowledge and System-Forming Character of Cognitive Activity*
 V.I. STOLJAROV, *Logico-Methodological Principles of Notion Unification in Modern Science*
 K.-H. STRECH, *Methodological Situations of Global Modelling Within the Context of the Theory of Science*
 R. STUPPOV, *Methodology of a Non-Traditional Type: Notes on Its Self-Negating Development*
 SU CHENGZHANG, LIU YUESHENG, *Pansystems Theory and Exploitation of Intelligence*
 SUN XIAOLI, *Model—the Core of the Methodology of Modern Science*
 A.V. SURIN, *Computerization and Its Impact on Scientific Research: the Main Stages of Development*
 V.M. SVIRIDENKO, *Ideal of Scientific Knowledge as a Form of Value-Oriented Consciousness in Science*
 Yu.B. TATARINOV, *Socio-Cognitive Principles of Evaluating Scientific Research Results*
 V.P. TIKHOMIROV, G.V. TELYATNIKOV, *Methodological Analysis of Correlation between Informatics and Cybernetics*
 I.S. TIMOFEYEV, *On the "Model" Conception of Science*
 S.A. TOLEDO, *On Relations between Classic and Stochastic Determinism*
 L. TONDL, *Explaining Technological Changes and New Paradigms in Technology Design*
 L. TONDL, *System of Knowledge as a Foundation of Knowledge-Based Systems*
 A.Y. TSOFNAS, *Theory-System Approach to Determining an Object and Criterion of Understanding*
 B. TUCHANSKA, *The Problem of Scientific Revolutions*
 I. TUDOESCU, *Symmetry and Unsymmetry as Laws of System Structure and Dynamics*
 J. TUHLENOV, *On the Forms of Theory's Influence on Scientific Thinking Style*
 V.S. TYUKHTIN, *On the Development of the Methodology of Empirical Knowledge*
 J. URBANIEC, *Contribution of Cosmology to the Understanding of the Empirical Method*
 P. VALEV, *Determinism and Reflection Theory*
 S. VASILEV-VASEV, *The Principle of Practice in the Structure of Scientific-Theoretical Knowledge*
 V.E. VOITSEHOVICH, *Modern Tendencies and Models of Mathematics Development*
 I.N. VORONTSOV, *On the Conceptual Basis of the Scientific Knowledge System Language*
 M. WAHBA, *Method of the Unity of Science*

- WANG DE-SHENG, The Method of Symmetry
 WANG SHUNYI, Prigogine as an Example: A Model of Cognitive Regulation in Scientific Discoveries
 WANG ZELI, The Theory of the Forms of Matter and the Hypothesis of the Space-Time Chain of Matter in the Universe
 S. WATANABE, When is Epistemic Relativity Inevitable?
 P. WEINGARTNER, G. SCHURZ, A New Approach to Verisimilitude
 R. WOJCICKI, Approximating to Truth
 WU TINGRUI, On the Reflection Method in Scientific Research
 A.V. YUREVICH, Philosophical Factors of Scientific Explication
 J. ZELENY, Dialectization: Remarks on Some Controversial Aspects
 J. ZEMAN, The Materialistic Theory of Motion and the Unity of the World
 ZENG XIAOXUAN, KOU SHIQI, Thinking of Image in Science and Technology
 ZHA RUQIANG, Four Great Achievements of Twentieth-Century Natural Sciences Have Enriched the Dialectics of Nature
 V.F. ZHURAVLYOV, Ordering Concepts and Terms in Branches of Science
 V.S. ZHURAVSKY, A.M. RUBANETZ, On the Regulative Role of Objectivity Principle in Philosophic-Methodological Reconstruction of Theoretical Cognition
 J. ZOMAN, The Materialistic Theory of Motion and the Unity of the World
 I.F. ZUBKOV, Order and Interconnection of Logical Forms of Theoretical Knowledge
 V.A. ZVIGLYANICH, Theoretical Structures of Knowledge in Non-Empiricist Versions of Western Methodology of Science

* * *

- A. SH. ABDULLAEV, O prirode mehanizma informatsionnykh protsessov
 S. ANGELOV, Metodologicheskiye problemy sovremennykh nauk
 V.I. BELOZERTSEV, Metodologiya issledovaniya nauki kak sotsialnogo instituta
 D.V. DZHOKHADZE, K kritike ponimaniya filosofii kak nauki nauk
 A.S. FOMIN, Mezhdistsiplinarnye svyazi i kompleksnye nauchnye issledovaniya
 N.V. GAIDUKOV, Modeli nauki e EVM
 V.A. GEROIMENKO, Neyavnoye nauchnoye znaniye i struktura subyektta
 A.V. KEZIN, Ideal nauchnosti v strukture poznavatelnoi deyatelnosti
 A. M. KORSHUNOV, Otsenka kak metodologicheskii printsip
 V.V. KOSTETSKII, Dialektika mifichnosti i nauchnosti nauki
 S.V. KOTINA, Rol problemnoi situatsii v razvitii nauchnogo znaniya
 V.A. KOVARSKII, Dialektika obratnykh svyazei i kooperativnye nelineinye modeli v osnovaniyakh sinergetiki
 A.K. KUDRIN, Ochevidnost i problema tipologii dokazatelstv
 I.I. LEIMAN, B.D. YAKOVLEV, Ob izmenenii granits metodologicheskoi bazy nauki
 V.I. LOKTIONOV, K metodologii issledovaniya znacheniya
 V.D. MAZUROV, Logika issledovaniya plokh formalizuyemykh tekhniko-ekonomicheskikh sistem sredstvami matematicheskogo modelirovaniya
 A.S. MAIDANOV, Tipologiya nauchnykh otkrytii
 A.S. MANASYAN, Subyektivnoye nachalo v strukture nauchnogo poznaniya i yego eliminatsiya
 V.V. MANTATOV, Problema teksta i yego ponimaniya
 S.F. MARTYNOVICH, Fakty nauchnogo znaniya v strukture istoricheskikh formatsii nauchno-poznavatelnoi deyatelnosti
 I.S. NARSKII, Metodologiya nauki i voskhozhdeniye ot abstraktnogo k kontretnomu
 V. Ye. NIKOFOROV, Metodologiya problem: osnovnye ponyatiya i protsedury

- PAVLOVA LILYANA, *Novaya forma nauchnoi kolektivnosti*
 Ya.K. REBANE, *Sotsialnaya pamyat i tsennostnaya orientatsiya nauchnogo poznaniya*
 V.M. ROZIN, *Metodologicheskii analiz traditsionnogo i netraditsionnogo proyektirovaniya*
 G.I. RUZAVIN, *Metodologiya ili logika nauchnogo otkrytiya*
 A.I. UVAROV, V.M. FIGUROVSKAYA, *Ob obshchem i spetsificheskom v metodologii
 tekhnicheskogo i sotsialnogo poznaniya*
 Z. ZASTAVKA, *Klassifikatsiya kak aktualnaya problema nauki*
 D.D. ZEBRILO, *Nauchnaya Shkola: Stimuly formirovaniya; kriterii, progress i regress*
 N.I. ZHUKOV, *Metodologicheskoye znachenie obshchei teorii sistem i kibernetiki v
 sovremennoi nauke*

Section 7. Foundation of Probability and Statistical Inference

- N.A. ALYOSHINA, *Probability Logic and Interpretations of Probability*
 S. AMBROSZKIEWICZ, *On Finite Random Sequences*
 P. BANKSTON, W. RUITENBURG, *Winning Theories and Probable Theories Using
 Trees of Finite Structures*
 C. BUTTASI, R. FESTA, *Generalized Carnapian Systems, Dirichlet Distributions and the
 Epistemological Problem of Optimality*
 N.C.A. DA COSTA, *Acceptance and Probability*
 J. EARMAN, *The Limits of Bayesian Inductivism: a Plea for Eliminative Induction*
 E. EELLS, *Inductive Probability and the Popper/Miller Argument*
 S. FAJARDO, *Game Relations between Adapted Stochastic Processes*
 A. FALK, *Using Bayes's Theorem Repeatedly to Corroborate Results*
 R. FESTA, *New Aspects of Carnap's Optimum Inductive Method*
 A.M. FINKELSTEIN, V.YA. KREINOVICH, *Impossibility of Hardly Probable Events:
 Physical Consequences*
 I.J. GOOD, *The Interface between Statistics and the Philosophy of Science (invited
 speaker)*
 I. HACKING, *Astronomical Improbability (invited speaker)*
 J. HUMBURG, *A Novel Axiom of Inductive Logic which Implies a Restriction of the
 Carnap Parameter*
 R. JEFFREY, *Probabilistic Epistemology: De Finetti's "Probabilismo" (1931)*
 A.S. KRAVETS, *On Epistemological Probability Types*
 I. KVART, *Causal Relevance*
 CHENG-HUNG LIN, *Hempel on Inductive Shortcoming in Craigian Method*
 L.G. MALINOVSKY, *A Substantial Basis for Statistical Inference*
 M. MARENCO, *Simplicity, Prior Probability and Inductive Arguments for Theism as a
 Cosmological Theory*
 I. NIINILUOTO, *Probable Approximate Truth*
 B.N. PIATNICZYN, *The Development and Formation of the Logics of Probabilities*
 J. VON PLATO, *Probability in Dynamical Systems (invited speaker)*
 J.H. SOBEL, *Two Partition-Theorems for a Causal Decision Theory*
 P. SUPPES, *Randomness and Determinism*
 M. VAN LAMBALGEN, *How Random is a Random Sequence?*
 V.N. VAPNIK, *Inductive Principles in Problems of Statistical Inference*

* * *

- O.E. DEMINA, *Tri napravleniya v sovremennoi induktivnoi logike*
 N.N. ENGVER, *Geometricheskii podkhod k resheniyu zadach ekstrapoyatsii i teoriiya
 statisticheskogo vyvoda*

A.V. KHAIDAROV, N.N. ENGVER, Metodologiya postroyeniya standarta istinnosti dlya udostovereniya pravilnosti vybora formy empiricheskikh zavisimostei

V.A. SVETLOV, O novom dokazatelstve K. Poppera nevozmozhnosti induktivnoi interpretatsii ischisleniya veroyatnostei

Section 8. Foundations of Physical Sciences

J.M. ABDULAEV, The Concept of Self-organization in Modern Physics from the Standpoint of Philosophical Conception of Reflection

R.R. ABDULLAEV, Relativity to Systems of Abstraction in the Development of the XX Century Cosmology

C. ABERG, Theoretical and Empirical Ingredients in Modern Physics. Philosophical Comments on the Developing Theory for Superstrings

N.T. ABRAMOVA, The Principle of Monism

I. APOSTOLOVA, The Methodological Notions in Physics — Dialogue Between Science and Culture

V.I. ARSHINOV, V.N. PERVUSHIN, Quantum Theory and Evolution of Modern Scientific Outlook

R.T.W. ARTHUR, Space as an Order of Situations: the Unappreciated Novelty of Leibniz's Relationalism

E.P. BALASHOV, V.V. SPIRIDONOV, A Functionally Structural Approach Concept and Systems Evolutionary Synthesis Methodology: Foundations and Problems

Yu. V. BALASHOV, Necessity and Change in Modern Cosmology

L.C. BARGELIOTES, The Perceiver and the Perceived in Contemporary Science

E.S. BOYKO, Methodological Premises of Analysis of the Structure and Genesis of the Theory of Non-linear Oscillations

V.P. BRANSKY, Heuristic Role of Philosophical Principles in the Formation of Fundamental Physical Theory

J. BUB, How to Solve the Measurement Problem of Quantum Mechanics?

M.E. BURGOS, Linearity Versus Symmetry: a Paradox?

S. BUTRYN, Dialectical Materialism and the Conception of the Quantum Origin of the Universe

J.T. CUSHING, A Naturalized, Socialized, but Highly Constrained Model of Scientific Change

G. DARVAS, Level Theories in Philosophy and Physics Exploring and Interpreting the Structure of Matter

K. Ch. DELOKAROV, Methodological Foundations of Regulative Principles of Physics

D.G. DIEKS, Special Relativity and the Flow of Time

H. DINGGUO, On the Ontological Basis for Quantum Theory

I.S. DOBRONRAVOVA, Problem of the Developing System Integrity in Contemporary Physics

A. DRAGO, Limit Principles in Theoretical Physics and Constructive Mathematics

V.N. DUBROVSKY, Basic Concepts of Space-Time in Physics

W. EBELING, Models of Physical and Biological Strategies in Evolution Processes

F.M. EFENDIEV, The Process Aspect of the Concept of Elementarity

M.A. ELYASHEVICH, T.S. PROTOKO, On the Problem on New Knowledge Formation in Physics

J. FAYE, The Problem of Realism in Quantum Physics

F. FELECAN, Contemporary Chemistry from Structural to Organizational Paradigm

- J. FENNEMA, "Give Me Where to Stand and I Will Move the Earth" — on the So-called Foundations of Physics
- N.A. GASANOV, R.O. KURBANOV, The Modern Physics and the Philosophy of the Ancient East
- K. GAVROGLU, "Translating" New Phenomena into Physical Problems
- V.A. GEROIMENKO, The Problem of Demarcation of the Physical Content and Mathematical Form of Theory
- V.S. GOTT, V.I. ZHOG, The Unity of the Linearity and Non-linearity of Physical Processes
- Y. GOUDAROULIS, Experimental Observation of New Phenomena and the Concept of Error
- L.M. GUTNER, Measurement in the Structure of Theoretical Relations
- P.A. HEELAN, Measurement as Interpretation
- D. HOME, N. MUKUNDA, The Bohr–Einstein Exchanges and Foundations of Quantum Mechanics: Sixty Years After the Complementarity Principle
- HU HAO, LOU HUI-SHIN, Some Views on Self-organization Theory from Philosophy
- G.M. IDLIS, Mathematical Principles of Science and Unity of Principle of Systems of Fundamental Structural Elements of Matter at All Successive Basic Levels of its Natural Self-organization
- R.J. ILIC, Analogie Scientifique à la Physique
- D.D. IVANENKO, On the Unified Theory
- M. JAMMER, The Concept of Distant Simultaneity in Classical Physics and in Relativistic Physics
- A. KAMLAN, Energy and Momentum Mass
- E.B. KATSOV, Mathematics and Physics in Topos "World"
- V.V. KAZYUTINSKY, The Anthropic Principle as a Philosophical Foundation of Cosmology
- U. Zh. KHAIDAROV, On Methodological Significance of Scientific Laws in the Development of Physical Knowledge
- V.N. KNYAZEYEV, The Principle of the Unity of Interactions as the Component of the Foundations of Modern Physics
- Y. KONGNIAN, New Dissertation on the Gravitation Outlook and God Outlook of Isaac Newton
- A.M. KRAVCHENKO, Normative-Value Approach to Substantiation of Physical Theories
- J. KRIKSTOPAITIS, The Physical System as a Unity
- Ju. I. KULAKOV, On the Unified Physical Image of Nature
- V.I. KUZNETSOV, M.S. BURGİN, Models, Laws and Principles in Physical Theories
- P.J. LAHTI, Uncertainty Relations — Formal Aspects and Interpretations
- B. LOEWER, D. ALBERT, Interpreting the Many-Worlds Interpretation
- Y. LOMSDAZE, T. LEPEKHINA, Alternate Physical Theories and Dialectics of Developing Their Scientific Language
- G. LUDWIG, An Axiomatic Basis as a Desired Form of a Physical Theory (invited speaker)
- E.M. MACKINNON, Classical Concepts and the Interpretation of Quantum Physics
- E.A. MAMCHUR, Informativeness of Theories as a Criterion of Scientific Progress
- K. MARTINAS, L. ROPOLYI, Aristotelian Thermodynamics
- MEI XIAOCHUN, Paradox of Checking Clocks
- MEI XIAOCHUN, The Re-interpretation of Quantum Mechanics
- N.V. MITSKIEVIC, Mathematics and Physical Reality
- Yu.B. MOLCHANOV, The General and Universal Character of Time
- B.Z. MOROZ, H-W. WIESBROCK, A Few Remarks Relating to the Notion of State-Vector in Quantum Mechanics

- F. MÜHLHÖLZER, Holes, Events and Symmetries in Space-Time Structures
 D. MUKHOPADHYAY, Physics in Galileo's Analysis of the Problem of Falling Bodies
 G.V. MYAKISHEV, Interpretation of Quantum Mechanics Before and After the Formulation of Quantum Statistics and Quantum Field Theory
 F.G. NAGASAKA, Rationality of Physics
 G. NERLICH On Learning from the Mistakes of Positivists (invited speaker)
 H.P. NOYES, D. MCGOVERAN, T. ETTER, M.J. MANTHEY, C. GEFWERT, A Paradigm for Discrete Physics?
 R.M. NUGAYEV, Origin and Resolution of Theory-Choice Situation in Theory of Gravity
 L.N. OAKLANDER, Metaphysics, Commonsense, and the Problem of Time
 V. PAMBUCCIAN, Four Variants for Eternity
 A.I. PANCHENKO, Foundations of Quantum Physics
 M. PAVICIC, Probabilistic Semantics for Quantum Logic
 V.P. PETKOV, The Flow of Time and the Conventionality of Simultaneity
 I. PITOWSKY, The Computational Intractability of the Generalized Bell Inequalities
 H.S. PLENDL, The Nature and Significance of Constants in the Physical Sciences
 A.P. POLIKAROV, Modular Conception of the Physical Theory
 D. POPA, I. BANSOITU, C. POPA, New Epistemologic Lineaments for the Interpretation of Quantum Mechanics
 A.R. POZNER, On Application of Complementarity as a Research Method
 B.Ya. PUGACH, Dialectics Correlation of the Observable and Unobservable in Scientific Cognition
 O.S. RASUMOVSKY, Variation of Axiomatics and the Unity Problem in Physics
 M. REDHEAD, Quantum Physics and the Identity of Indiscernibles
 M.C. ROBINSON, Is Dialectical Materialism Relevant to Modern Physics?
 L. ROPOLYI, Dualities and Individual Dynamics in Classical Physics
 U. RÖSEBERG, Evolution and Physics
 N.M. ROZHENKO, Quantum-logical Justification of the Aristotle Qualitative Physics
 M. SABITOV, On the Role of Bohr's Conformity Principle in Quantum Mechanics Formation
 Yu.V. SACHKOV, Physics-Biology Interrelationship: Towards a New Paradigm
 S. SEMCZUK, On Relation Between Empirical and Abstract Objects: a Short Study in the Problem of Mathematization
 A.L. SIMANOV, The Epistemological Status of Axiomatics in Physics
 S.P. SITKO, A Physical Criterion of Stable Integrity of Self-Organizing Systems
 N.M. SRETENOVA, The Ancient Chinese Model of the World (Taoism) and Modern Physics
 L. STANIS, Methodological Foundations of One Revolution in Physics
 A. STEFANOV, V. PETKOV, What is the Quantum Mechanical Object?
 I. STEIN, Past and Future in Quantum Mechanics
 V.S. STEPIN, Scientific Revolutions and Potentially Feasible Trends in the Historical Development of Physics
 M. STOECKLER, The Role of Symmetries in Quantum Field Theory
 A.T. STRIGACHEV, Fundamental Problems of Quantum Physics as Viewed from the Symmetry Principle (for the Purpose of Methodological Principles System)
 P. SZEGEDI, Indeterminism in Quantum Mechanics
 P.C.L. TANG, N. BASAFA, Complementarity in Nonequilibrium Thermodynamics
 M. TEMPCZYK, Black Holes and Unity of Physics
 F. THIEFFINE, D. EVRARD, Logic, Probability and Models: Hidden Variables and Semantical Constraints in Quantum Mechanics

- V.G. TOROSIAN, The Style of Thinking as a Component of the Foundations of the Physical Knowledge
 A.V. TYAGLO, I.Z. TZECHMISTRO, The Idea of Integrity as a Necessary Element of a Conceptual Foundation of Quantum Theory
 A.A. TYAPKIN, On Irreversibility in Statistical Physics
 G. VERSTRAETEN, To a Microscopic Foundation of Entropy?
 R.A. VIHALEM, On the Problem of Methodological Identification of Chemistry with Physics
 S.V. VONSOVSKY, V.I. KORYUKIN, G.G. TALUTZ, To a Critical Analysis of the Foundation and Developmental Vestas of Modern Physics
 T.P. VORONINA, Features of Theory Development in Modern Physics
 WU YUAN-FU, The Philosophy Problem in the Model of Inflationary Universe
 P. ZEIDLER, The Heuristic Role of Mathematics in Physical Sciences
 S.N. ZHAROV, Inceptive Abstract Objects as a Channel of Socio-Cultural Determination of Science
 G.B. ZHDANOV, Conceptual Spheres of Science and Their Interaction

* * *

- I.A. AKCHURIN, Fundamentalnye fizicheskiye teorii i topozy
 A.V. KATSURA, Problema evolyutsionnoi fiziki
 Sh.Yu. LOMSAZDE, Nils Bor, Albert Einstein n problema fizicheskoi realnosti
 A.M. MOSTEPANENKO, Problema determinatsii sushchestvovaniya v fizicheskom poznanii
 ANTAL MYULLER, Problemy obyektivnosti i adekvatnosti poznaniya v fizike
 P.I. PIMENOV, Samostoyatelnoye mesto "gladkosti" v sisteme "diskretnoye-nepriyemnoye" v teorii prostranstva-vremeni
 A.A. VASILCHENKO, Otrazheniye i proyavleniye informatsionnoi prichinnosti v fiziko-khimicheskikh i geologicheskikh protsessakh

Section 9. Foundations of Biological Sciences

- A.I. ALYOSHIN, On the Methodological Status of the Theory of Evolution
 L. BELKA, K. ZEMEK, The Principle of Uniformitarianism in the Past and Present Evolutionary Thinking
 F. BONSAK, A Causal Interpretation of Biological Teleology
 D. BUICAN, Philosophy of History of Biology: Revolution of Evolution
 J.D. COLLIER, Use of the Entropy Concept in Evolutionary Biology
 N.P. DEPENCHUK, On the Methodological Foundations of Biological Knowledge Integration
 M. FOREMAN, The Methodology of Taxonomy
 FU JIEQING, Tendency of Alterations of Scientific Methods in Biomedical Sciences of Recent 100 Years and Two Models of Biomedical Discoveries
 J.M. GALL, The Conceptual Bases of the Synthesis of Population Ecology, Genetics and of the Evolutionary Theory
 E.V. GIRUSOV, The Necessity of the Formation of Sciences of Biosphere Cycle
 M.A. GOLUBETZ, Methodological Foundations, Structure and Problems of Modern Ecology
 HE YU-LIANG, Principle on Disordered State Evolved Spontaneously into Ordered State of Nature
 M.O. IBODOV, "Tendency" Concept: Its Role in Organic Evolution Cognition

- J. JANKO, Physico-Chemical Biology, Theoretical Biology and Unity of the Science
- ZHANG JI, The Ontology of Human Life
- Z.V. KAGANOVA, Anti-essentialism and Biology
- T.V. KARSAEVSKAJA, Humanitarian Values as a Foundation for Studying the Dialectics of a Human Life-Cycle
- L.P. KIYASHCHENKO, On the Problem of the Wholeness of the Object of Socio-Ecological Knowledge
- E.I. KOLCHINSKIJ, The Problem of "Evolution of Evolution" and the Global Evolutionism
- V.A. KRASSILOV, Singular and General in the Methodology of Natural Sciences
- A.Z. KUKARKIN, Genetics and Psychology: Theoretical Interaction or Methodological Alternative
- S.V. KUPTSOVA, The Study of Anthropogenesis and its Peculiarities
- V.A. KUTYREV, Methodological Preconditions for Formation of Coevolution Systems
- V.V. LEONOVICH, The Level of Biological Integration and Organism as a Touchstone of Biological Investigation
- R.C. LOOIJEN, Emergence and Reduction on Biology
- V.A. LOS, Methodological Problems of Socio-Ecological Studies
- R. LÖTHER, Evolution — Matter of Fact or Metaphysical Idea? (invited speaker)
- A.S. MAMZIN, On Interrelationship of Organization and Development of Living Systems
- G.M. MARTYNENKO, Philosophical and Methodological Aspects of the Problem of Biological Systems Autonomy
- M.N. MATVEYEV, Philosophical Problems of Cognition of Prebiological Evolution
- E. MIRZOJAN, The Evolutionary Synthesis: Methodological Aspect
- S.D. MITCHELL, Are Sociobiological Adaption Explanations Legitimate?
- J. MLIKOVSKY, Fundamental Units and Reductionism in Evolutionary Biology
- J. MOSTERIN, The Role of Recursive Definitions in Biological Ontology
- S.A. NICKOLSKI, On the Problem of Coordination of Biological and Social Knowledge in the Study of Human Behaviour
- L. NISSEN, Three Ways of Eliminating Mind from Teleology
- Ju. I. NOVOZHENOV, Evolution under the Influence of Culture
- B.V. PREOBRAZHENSKY, System Approach in Modern Biology
- A. RAY, Prediction of Kinship and Inbreeding from Populations of India
- M.H. REMMEL, On two Fundamental Explanatory Systems of Biological Development: Paradigms of Baer and Darwin
- N. ROLL-HANSEN, The Practice Criterion and the Rise of Lysenkoism
- S. SANDER, Evolution — Punctualism, Gradualism or Neutralism?
- R.T. SARSENOV, "Man-Land" Relations as One of the Principal Foundations of Modern Scientific Weltanschauung
- M.L. SHAMES, The Arche: A Model of Cultural Evolution
- A.T. SHATALOV, On the Directions of Evolutionary-Historical Progress and Individual Development of Human Organism
- B.A. SIGMON, The Influence of J.T. Robinson on the Development of Thought in Human Paleontology
- E.R. SOBER, Evolutionary Altruism and Psychological Egoism (invited speaker)
- A.A. SOZINOV, Philosophical Problems of Modern Genetics
- T. SUTT, The Synthetic Theory of Evolution and the Anthropic Principle
- P. TAYLOR, The Strategy of Model Building in Ecology, Revisited
- S.V. TCHERKASSOV, Logical and Methodological Principles of Experimental Studies in Clinics
- P.D. TISCHENKO, Marxist Concept of Man's Nature and a Possibility of a New Research Perspective in Human Biology

- L.J. VEVERKA, Philosophic Category of Quality and the Notion "Biological Species"
 K. WENIG, Evolution and Progression
 YE YONGZAI, The Significance of Bio-Holographic Law in Philosophy and Scientific Methodology
 B. ZEIDE, Methodology of Biological Modeling
 A.I. ZELENKOV, P.A. VODOPIANOV, Bifunctional Status of Cultural Traditions in the Development of Ecology as a Science
 YINGQING ZHANG, The Bio-Holographic Law, Eciwo (the Embryo Containing the Information of the whole organism), the Eciwo Theory, and Eciwo Biology
 A.T. ZUB, On the Problem of Unity of Levels in Theoretical Description of the Progress of Biological Evolution

* * *

- V.D. BELYAKOV, G.D. KAMINSKII, Otrazheniye i upravleniye v samoorganizatsii biologicheskikh sistem
 A.N. CHUMAKOV, Novoye myshleniye kak faktor resheniya globalnykh problem
 N.I. DEPENDCHUK, K voprosu o metodologicheskikh osnovaniyakh integratsii biologicheskogo znaniya
 K.M. HAILOV, Istoricheskaya i kontseptualnaya vzaimosvyaz biologicheskoi, bioekologicheskoi i biogeokhimicheskoi modelei zemnoi prirody
 K.N. HON, O vzaimodeistvii biologicheskikh sposobov myshleniya
 R.S. KARPINSKAYA, Filosofiya i problemy sinteza biologicheskogo znaniya
 A.A. KOROLKOV, Metodologiya teorii normy v mediko-biologicheskikh naukakh
 I.Ya. LEVYASHCH, Predmet sotsialnoi ekologii i biologicheskoye nauki
 Ye.N. SHATALIN, Modelnyi fetishizm v metodologii eksperimentalnoi genetiki
 V.I. VASILENKO, Ekologiya i antropotsentrizm
 K. ZEMEK, V. NOVOTNYI, V. LEONOVICH, Nekotorye printsipialnye predposylki razvitiya sistemno-evolyutsionnogo podkhoda v biologii

Section 10. Foundations of Psychology and Cognitive Science

- H.B. ANDERSEN, The Notion of Tacit Knowledge in Philosophy and Cognitive Science
 E. ANDREEWSKY, V. ROSENTHAL, The Interplay between Neurolinguistics and Psycholinguistics: a Methodological Tool for Cognitive Science
 T. BALOGH, On Conception of "Equilibrium"
 M.C. BARLIBA, The Mechanism of "Double Idealisation" of Information in Contemporary Communication
 P.G. BELKIN, Methodological Problems in Studying the Sociopsychological Adaption
 I.A. BESKOVA, Subconsciousness: Some Aspects of Functioning
 N. BLOCK, Functional Role and Truth Conditions
 M.S. BOGNER, A Case for Unipolar Dimensions in Psychological Testing and Knowledge Representation
 A.A. BRUDNY, Understanding as a Subject of Matter of Psychological Research
 A.V. BRUSHLINSKY, Concerning One of the Specific Foundations of Psychology
 M. BUNGE, The Place of Psychology in the System of Knowledge
 J.M. CHAMORRO, Teorias de la Mente y Psicologia Cientifica
 CHEN TI DIAN, LI HONG, On the Combination of Modular Thought with Structural Thought

- E.V. CHERNOSVITOV, Consciousness in the Structure of Selfconsciousness: Towards the Analysis of the Objective Reality
- S.V. CHESNOKOV, Humanitarian Measurements as a Basis for Obtaining Humanitarian Knowledge
- A. CLARK, Cognitive Science and the Dogma of Implementation-Neutrality
- M. COLBERG, M.A. NESTER, The Use of Illogical Biases in Psychometrics
- A. DANAILOV, C. TÖGEL, Reflection and the Theory of Evolution
- E.N. EMELYANOV, I.G. POSTOLENKO, Methodological Attitudes in the Researcher's Activities and Problems of Interaction in a Research Team
- E. ERWIN, Psychoanalysis; Clinical vs. Experimental Testing
- M.A. GELASHVILI, The Problem of Functional Analysis of Attitude
- V.D. GLEZER, Vision and Mind (invited speaker)
- J. HUMPHRIES, Artificial Intelligence and Human Mental States
- I.V. IMEDADZE, On System-Formative Factor of Behaviour
- A.M. IVANITSKY, Certain Principles of the Organization of Cerebral Functions Underlying Psychics
- Ch.A. IZMAILOV, Color Specification of Emotions
- B. KEANEY, R. DOUGLAS, A Critical Appraisal of Two Theories of Brain and Mind
- O.G. KOGAN, V.A. MINENKOV, The Problem of Subject-Object Relationship in Deontology
- D.H. KRANTZ, The Role of Axiomatization in Behavioral Science Theories (invited speaker)
- N. LACHARITE, Aspects Constructivistes de L'Approche Informationnelle en Théorie de la Connaissance
- E. LEPORE, A Dialogue Between Two Kinds of Realists
- J. LINHART, Theory of Thinking and Models of Creativity
- A. LISSOWSKI, Two-Dimensional Contiguity Scaling
- J. MACNAMARA, Logic, Psychology and Proper Names
- N.S. MANSUROV, The Principle of Development and the Contemporary Problem of Psychology
- A. MARRAS, Mental Images and the Frame Problem in Artificial Intelligence
- N.B. MICHAILOVA, Methodological Basis of the Problem of Abilities
- I.A. MORARU, M.I. MORARU, A Cybernetic Model of Moral Norms Internalisation and of Moral Behaviour Moulding
- V.F. MORGUN, The Type of Personality: The Unity of Monism and Multimetrisism
- NG TAI-KEE, Existence and Types of Consciousness
- S.A. PEDERSEN, Mathematical Models in Cognitive Science
- O.F. POTEMKINA, Methodological Tendencies in the Studies of Mental Reflection
- K. PSTRUZINA, Principles of Dialectic Theory of Thinking
- A.P. RASTIGEYEV, Methodological Problems of Studying Social Adaption of Personality
- J.D. RINGEN, Freedom from Stimulus Control: Aristotelian, Cartesian and Darwinian Perspectives
- V.S. ROTENBERG, The Role of the Two Strategies of Thinking in the Process of Scientific Cognition
- C. SCHÜES, Implications of Wittgenstein's Philosophy for Psychology
- B. SHANON, The Limitations of the Representational View of Mind
- A. V. SOKOLOV, Foundations of Psychoanalysis: The Problem of Basic Principles in Freudian Theory
- E.N. SOKOLOV, Psychophysiology and Artificial Intelligence
- N. STANOULOV, Decision Making as a Basic Paradigm of Human Thinking
- C. SWIFT, Logical Inconsistencies in Mathematical Models of Neural Systems

- A.Yu. TEREKHINA, Geometric Model of Knowledge Structure
 M. THOMAS, The Conflict Between Methodology and Ethics in Psychological Research
 B. UMBAUGH, The Question of Normative Inferential Pluralism: Why Worry?
 M.M. USMANOV, On the Gnoseological Functions of Communication
 V.F. VENDA, The Laws of Mutual Adaptation, Their System and Psychological Aspects
 K.V. WILKES, Of Mice and Men: The Comparative Assumption on Psychology
 V.A. YAKOVLEV, Le Principe Structurel de Classification des Programmes de Recherches en Psychologie de la Pensée
 A.V. YUREVICH, Psychological Study: Problem of Organization
 ZHOU YI ZHENG, Intuitive Thinking and Illogical Method in Scientific Creation

* * *

- G.G. ARAKELOV, Neironnaya organizatsiya v determinatsii povedeniya prostykh sistem
 N.N. DANILOVA, Differentsialnaya psikhofiziologiya kognitivnykh protsessov
 E.A. GOLUBEVA, Psikhofiziologiya poznavatelnykh protsessov
 V.I. KAPRAN, S.A. SHLYKOV, Pertseptivnye protsessy
 A.A. LEONTYEV, K metodologii issledovaniya rannikh etapov istoricheskogo razvitiya psikhiki y cheloveka
 A.A. PONUKALIN, Psikhologiya poznavatelnoi deyatelnosti
 S.M. SHALYUTIN, O nekotorykh soderzhatelnykh interpretatsiyakh teorem Makkaloka i Pittsa o neironnykh setyakh
 V.G. SHVYRKOV, Sistemno-evolyutsionnyi podkhod k analizu mozgovykh protsessov
 Ye.N. SOKOLOV, Psikhofiziologiya i iskusstvennyi intellekt
 O.K. TIKHOMIROV, Informatika i psikhologiya
 B.M. VELICHKOVSKII, Funktsionalnaya struktura poznavatelnykh protsessov: smena paradigmy

Section 11. Foundations of the Social Sciences

- G. ANDRASSY, What is "Notional Grasping" in Marx?
 T.V. ARTEMYEVA, Special Functional Features of the Historical Types of the Social Utopianism
 C. BICCHIERI, Strategic Behavior and Counterfactuals
 J. BOULAD-AYOUB, Thesis 7 on Feuerbach Revisited: Towards a Reevaluation of Ideological Process and Its Societal Efficiency
 L. BOVENS, An Essay on Two-Person Bargaining Theory
 H.A. BROWN, The Supervenience of the Social on the Individualistic
 E. BRUCKNER, A. SCHARNHORST, Stochastic Models and the Complexity Problem in the Social Sciences: An Example from the Theory of Self-organization and Evolution
 Y.F. BULAYEV, Man as Subject and Object of Historical Progress
 A. BURACAS, Metaeconomic Institutionalization of Conceptual Criteria of Social Development
 C. BURRICHTER, Stellung und Funktion der Sozialwissenschaften in den postindustriellen Gesellschaften
 V. CECHAK, The Problems of the Conception of Fact and Facticity in Social Sciences
 L.J. COHEN, A Note on the Evolutionary Theory of Software Development
 F. COLLIN, Natural Kind Terms and Explanation of Human Action
 B. DAJKA, Understanding Explanation
 G. DE MEUR, Snapshots on Mathematical Thinking in Political Science

- A. DIAS DE CARVALHO, Towards a New Epistemological Model for the Sociology of Education
- A. DRAGO, An Analogy between Marx' Theory and S. Carnot's Thermodynamics
- M.A. DRYGIN, On the Foundations of Methodological Reconstruction in the Development of Social Sciences
- G.G. DYUMENTON, Reconstruction of Cognitive Processes in Science: Socio-Psychological and Organizational Aspects
- J. ELSTER, Rationality and Social Norms (invited speaker)
- V.G. FEDOTOVA, Methodological Aspects of Philosophical Analyses of a Man
- D.W. FELDER, Logical Models of Conflict
- A.M. GENDIN, Social Prognostics and Its Methodological Functions
- N. GENOV, Towards a Synthetic Approach to Laying the Foundations of Sociological Cognition
- D. GINEV, Dialogical Type of Scientific Rationality—Towards a Methodology of Humanities
- J. GÖTSCHL, Elements of Self-Organisation: New Foundations of Social Sciences?
- A. GRZEGORCZYK, Some Theses on the Foundation of Ethics
- P.S. GUREVITCH, Rational as a Mean of Vital Orientation of a Man
- B. HAMMINGA, Labour as a Utility
- H. HATTORI, On Understanding an Alien Culture
- M.V. IORDAN, On the Problem of Subordination of Logical Aspects of Social-Historical Evolution
- M.C.W. JANSSEN, Utilistic Reduction of the Macroeconomic Consumption Function
- V.Zh. KELLE, Social Knowledge: Criteria of Its Scientific Nature
- A.A. KHAGUROV, Methodological Problems of Social Experiment and Social Cognition Interaction
- N.I. KIYASCHENKO, Aesthetics and the Methodology of Investigating Aesthetic Conscience
- K. KOEV, Everyday Life. The Invisible Horizons of the Self-Evident
- N.N. KOZLOVA, Cognitive-Mediating Function of Common Consciousness in Social Cognition
- S.E. KRAPIVENSKY, Cognition of Social Regularities through Historical Parallels
- T. KUNCA, On the Theory of the Social-Scientific Law
- M. KUOKKANEN, A Generalization of a Theory on Conflicts of Role Expectations
- T.F. KUZNETSOVA, The Problem of Correlation of Internalism and Externalism in Social Cognition
- M. LAGUEUX, Instrumentalism and "Constructive Empiricism" in Economics
- LEI DERSEN, On the Simultaneity of Discovery and Invention
- V.F. LEVICHEVA, Methodological Basis of Social Economic Knowledge (Why the Economics of Spirit is Possible?)
- E.N. LOONE, Sequential Dependencies in History and in Other Social Sciences
- B. MACKINNON, The Economic Value of a Life
- N.G. MAGOMEDOV, The Ideal of Explanation in Social Knowledge
- E.S. MARKARIAN, On the Interrelation between Sociology and Culturology of Science: Introducing the Problem
- J.E. MCCLELLAN, Proposals toward a Formal Definition of Social Class
- K. MELLOS, Anthropological Assumptions of Ecological Theories
- R. MIGUELEZ, Le Statut Epistémologique de la Nouveauté en Sciences Sociales
- Y. MINKOV, On the Value Approach to the Social Problems of the Scientific-Technological Revolution
- R. NADEAU, Hayek and the Methodological Peculiarities of Social Sciences
- P. NEL, The Limits of Marxist Functional Explanations: A Test Case

- V.J.A. NOVAK, Objective Ethics and the Principle of Sociogenesis
 H. NURMI, Intuitions Concerning Best Decisions in Collective Voting Bodies
 A.S. PANARIN, On Two Paradigms of Contemporary Bourgeois Social Thought
 I. PÖRN, On the Nature of a Social Order (invited speaker)
 A.-C. PREDA, The Structure of Social Theories
 R. RICHTER, The Dominance Argument and Two Paradoxes of Rational Choice
 P.H. ROSSEL, The Empirical Study of Moral Conceptions. A Method
 R. SASSOWER, The Quest for Scientific Respectability: Classical Economics in England
 N.Kh. SATDINOVA, Towards the Evolutionary Justification of Ethics
 C. SAVARY, An Interpretation of Popper's Third World Thesis
 M. SCHABAS, Making Sense of Sen
 I. SELEZNEV, On the Dialectics of Parametres of Creatively Transforming Activity
 Z. STAIKOV, A. KARDASHEVA-ULYKHINA, Materiality of the Social and Social Time
 L. STRIEBING, Computerisierung—eine Herausforderung für die Gesellschaftsphilosophie
 G. SZILAGYI, Aristotle's Conception of Economy as a Contribution to the Foundation of Economic Thinking
 T. TAKAMATSU, Social Aspect of Many-Valued Deontic Functors
 O. TENZER, On Research Method Transfer between Natural and Social Sciences (Example of Tomograph)
 R. TUOMELA, Supervenience as a Central Metaphysical Notion within the Social Realm
 D.J. VALEYEV, The Definition of Morality as One of the Most Important Fundamentals of Science of Ethics
 N.K. VASSILEV, Dialectical Mechanism of the Transition to Communist Society
 A.K. VOSKRESENSKY, Specifics of Social Sciences and Informatics
 R. WARE, Why Theorize about Groups? Why Not?
 S. ZAMAGNI, On the Epistemological Status of Economic Laws
 N.S. ZLOBIN, The Active (Subjective) Character of the Development of Science and the Reasons for Integrating Natural and Social Sciences

* * *

- Z.I. FAINBURG, G.P. KOZLOVA, Vychleneniye obyektivnogo v subyektivnom kak kriticheskii punkt prikladnogo sotsialnogo poznaniya
 G.N. GUMNITSKII, Etika i metodologiya
 V.Ye. KEMEROV, Gumanizm i ratsionalnost sovremenno go obschestvoznaniya
 S.I. KORDON, Dvoistvennyi kharakter yazyka ekonomicheskoi nauki
 V.M. SEMENOV, O metodologicheskome znachenii kategorii "protivorechiye" v issledovanii dinamiki razvitiya natsionalnykh othoshenii pri sotsializme
 A.A. STARCHENKO, Logicheskkiye osnovy pravovoi germenevtiki
 R. VALENCHIK, Iskhodnoye teoretiko-metodologicheskoye polo zheniye, yego struktura i funktsii
 O.N. ZHEMANOV, Operatsionalnyi i teoreticheskii urovni osnovanii obschestvennykh nauk

Section 12. Foundations of Linguistics

- N.L. ABRAHAMIAN, The Object and Method of Linguistics
 V.M. ALPATOV, On Intuitive and Research Approaches to Language Studying

- V. ANASTASSOV, Continuity and Discontinuity in Scientific Knowledge of Language
 A. BARULIN, On the Construction of the Scheme of the Language Integral Model
 I. BELLERT, M. ZAWADOWSKI, Quasi-Order of Quantification Structure Instead of Scope Method
 I.A. BESKOVA, Perspectives and Restrictions of Logically-Semantic Means in Natural Language Analysis
 A.V. BESSONOV, Truth Within the Language and Substitutional Quantification
 L.F. BONDARENKO, Structure and Functioning of Political Culture in the System of Historical Social Relations
 T.V. BULYGINA, A.D. SHMELEV, Non-Empirical Criteria for the Evaluation of Linguistic Description
 W.S. CRODDY, Language and Illocutionary Speech Acts
 S. CUSHING, A Class of Tri-valent Quantifiers from Natural Language
 I. DÖLLING, Natural-Language Negation-Aspects in the Relationship Between Logic and Linguistic Semantics
 A. DORODNYKH, A. MARTYNYUK, Critical Analysis of Philosophy and Empirical Results of Feminist Linguistics in the West
 B.D. DYANKOV, The Conception of Polysemantic Structure of Natural Languages and the Related Heuristics
 G. FORRAI, The Role of a Metaphor in the Birth of Generative Grammar
 J. GEURTS, T. BRINK, Reference and Convergence
 B.Yu. GORODETSKY, Linguistics and Computerization of Human Activities
 G. GRINENKO, Logic, Linguistics and Demonstratives
 J. HINTIKKA, G. SANDU, Informational Independence as a Semantical Phenomenon (invited speakers)
 V.V. IVANOV, On Principles of Constructing Linguistic Metatheory (invited speaker)
 M. JOHNSON, Grammar as Logic, Parsing as Deduction
 A.N. KOCHERGIN, O.A. DONSKYKH, The Peculiarities of the Reflexion on the Foundations of Linguistic Theory
 J. KOSTER, How Natural is Natural Language? (invited speaker)
 N.Z. KOTELOVA, Possibilities of Using the Data of Natural Language in the Processes of Scientific Cognition (to the Basing of Linguoheuristic)
 M.I. KOZLOV, B.I. PRUZHININ, On Gnoseologic Status of Analytical Procedures in Linguistic Semantics
 S.N. KUZNETSOV, Ontological Qualities of Language and Possibilities of Their Verification
 M. LAKOVA, On a Quantification Model of Sentence Semantics
 F. LATRAVERSE, The Role of Context in a Linguistic Theory
 FU-TSENG LIU, Some Evidences for the Annoyance of the Occurrence of Tense in Verbs
 B.P. MOSS, Unnatural Languages
 D. NESHER, Language Rules and the Evolution of Human Cognitive Behavior: Calculus, Language-Game and Beyond
 S.E. NIKITINA, The Semantic Structure of Scientific Knowledge: Linguistic Aspects
 E.V. PADUCHEVA, On Semantic Contents of Identity Statements in Natural Language
 V. PANZOVA, Logic Aspects of the Language
 R.J. PAVILIONIS, Understanding Linguistic and Nonlinguistic Texts: Intentionality, Intensionality and Indexicality
 V.M. PAVLOV, The Object of Linguistics and the Subject of Linguistic Research
 F.J. PELLETIER, Logics for Realistic Vagueness
 V.B. PERICLIEV, Heuristics in a Linguistic Investigation
 P.L. PETERSON, Real Logic Explanation

- V. PETROV, *Metaphors and Natural Language Understanding*
 G.G. POCHOPTSOV, *Functionalism in Syntactic Description*
 O.G. POCHOPTSOV, *Role Analysis as a Method of Linguistic Investigation*
 I.V. POLJAKOV, *Semantics and Classification of Paradigms for the Philosophy of Language*
 M.N. PRAVDIN, *Semiotics and Dialectics*
 B.-O. QVARNSTRÖM, *On the Semantics of Natural Kind Words*
 Z.M. SHALYAPINA, *A Model of Human Linguistic Competence as a Unified Basis for a Set of Fundamentally Dissimilar Models of Linguistic Performance*
 S. SHIMIDZU, *A Logical Approach to the Parsing of Natural Language*
 M.A. SIVERTSEV, *Typological Operator in Natural Logic of Typological Reasoning*
 K. SOLT, *On Language and Picture*
 Yu.S. STEPANOV, *On the So-Called "Logic of Language"*
 P. SWIGGERS, *A Foundational Problem in Linguistic Description: Rule Transparency*
 V.N. TELIYA, *Metaphor as a Manifestation of the Anthropocentric Principle in Natural Language*
 M. UJVARI, *Why Kantian Transcendental Philosophy Can Not Provide for an Analysis of Language?*
 D. VANDERVEKEN, *A Logical Theory of Non Literal Meaning*
 S.A. VASILIEV, *On the Problem of the Text Sense Growth in the Course of Its Social-Historical Functioning*
 N.B. VYATKINA, *Language and Ontology*
 V.K. ZHURAVLEV, *In Search of 20th Century Linguistics Basis*
 L.V. ZLATOUSTOVA, A.M. EGOROV, I.A. RASKIN, *Some Restrictions on Machine Modelling of Human Communication, Language and Thought*
 V.A. ZVEGINCEV, *Status of the Science of Language in the Modern World*

* * *

- N.D. ARUTYUNOVA, *K osnovaniyam lingvisticheskoi teorii*
 A.P. BARULIN, *K postroyeniyu skhemy integralnoi modeli yazyka*
 V.I. GERASIMOV, *K izucheniyu bazy znaniy, ispolzuyemykh v protsessakh yazykovogo funktsionirovaniya*
 E.B. KLEVAKINA, *O tipakh ratsionalnosti epistemicheskikh modalnostei*
 Ye. S. KUBRYAKOVA, *Osnovaniya sovremennoi lingvistiki i chelovecheskii faktor*
 V.I. POSTOPALOVA, *Osnovaniya lingvistiki v aspekte vnutrinauchnoi refleksii*
 V.M. SERGEYEV, *O graficheskoi tekhnike polucheniya vyvodov v ramkakh logiki yestestvenno-yazykovykh rassuzhdenii*
 A.M. SHAKHAROVICH, G.V. IVANOVA, *Metodologicheskkiye osnovaniya lingvisticheskogo obespecheniya sistem iskusstvennogo intellekta*
 E.M. VOLF, *Otsenka kak vid modalnosti*

Section 13. History of Logic, Methodology and Philosophy of Science

- WAZIR HASAN ABDI, HASAN MANZIL, *A Methodology of Science Prevalent in India During the Medieval Period*
 I.I. ALEKSEIEVA, *The Logical Structure of Aristotle's "Dialectical Conversation"*
 I.H. ANELLIS, *Russell's Problems with the Calculus*
 M. ASTROH, *Abaelard on Modalities De Sensu and De Rebus*
 R. AVRAMOVA, *Against Traditional Rationality*

- N.S. AVTONOMOVA, Problem of Rationality in the Light of Social Determination of Cognition
- I. BANSOIU, On the Concept of Irreducible Antinomy
- A.G. BARABASHEV, Z.A. SOKULER, Two Directions in the Contemporary Philosophy of Mathematics
- H. BARREAU, Lazare Carnot's Adhesion to the Leibnizian Conception of the Calculus
- H.S. BATISCEV, Humanitarianization and Axiologization of Knowledge as a Tendency in History of Cognition
- V.A. BAZHANOV, The Making and Development of N.A. Vasiliev Logical Ideas
- R.S. BHATNAGAR, Scientific Inquiry and Human Disposition
- BIAN CHUNYAN, In Combining with Practice to Study the Dialectics of Nature
- L.G. BIRYUKOVA, On the History of Genetic Method in Development of Deductive Sciences
- A. BOBOC, Aporetics and Its Place in Aristotle's Theory of Science
- L.A. BOBROVA, Reestimation of Frege's Ideas in Logic of Science
- A.G. BOTEZ, The Outlook of the "Stylistic Field" in the History of the Philosophy of Science
- T.G. BUCHER, A Fatal Presupposition of the Ontological Argument
- H. BURKHARDT, Ganzes–Teil Relation in Logik, Mathematik und Philosophie
- B.V. BURYUKOV, Mathematics and Logic: a Problem of their Relationship
- M.F. BYKOVA, Hegel's Methodology of Logicism: Real Content and Problems
- V.S. CHERNYAK, Science and Non-Science: The Criterion of Demarcation
- A. DRAGO, Money, Objective Scientific Knowledge and Social Consciousness
- F. DUCHESNEAU, Leibnitz and the Philosophical Analysis of Science (invited speaker)
- J. ETCEMENDY, Tarski, Semantics and the Liar
- G. EVEN-GRANBOULAN, Logique et Economie: les Limites de la Rationalité Economique
- FAN DAINIAN, The Development of Philosophy of Science in China in the First Half of the 20th Century
- B.I. FEDOROV, On the Relations of Equality and Inequality in the Deductive Theory of B. Bolzano
- M. FEHER, The 17th Century Crossroads of the Mathematization of Nature
- HANTING FENG, Philosophical Understanding of the Controversy on Quantum Theory
- H. FESTINI, Antirealism/Realism of Wittgenstein's and Hintikka's Language-Game Idea
- Z.Y. FILER, Historical and Philosophical Problems of the Continuous and Discrete in Mathematics
- MENACHEM FISCH, From the Point of View of Problems: Rereading Whewell
- M. FLONTA, Intertheoretical Correspondence Relation: the Standard Realist and the Copenhagen School Interpretations
- J. GASSER, Allegations of Premise Smuggling and their Historical Significance
- A. GILLIAM, Philosophical and Methodological Problems of the Anthropology of Development
- G. GIMPL, Logik Versus Physikalismus? Zu Wittgensteins Begründung der Logik und Naturwissenschaft
- E. GLAS, Are There Kuhnian Revolutions in Mathematics?
- M.S. GLASMAN, Theoretical and Aesthetical Principles in Natural Scientific Thinking of New Time
- N.B. GOETHE, The Kantian Roots of Frege's Thought
- GONG YUSHI, The History of the Studies of Dialectics of Nature in China
- P.I. GRADINAROV, The Quodammodo Logic of Indian
- A.F. GRIAZNOV, The Problem of "Conceptual Necessity" in L. Wittgenstein's Works
- A.G. GROSS, On the Shoulders of Giants: the Rhetoric of Light in the 17th Century

- C.D. GRUENDER, Scientific Method and the Lesson of Galileo
- L. HAAPARANTA, Analysis as the Method of Logical Discovery: Some Remarks on Frege and Husserl
- G. HEINZMANN, Zur Ausbildung pragmatischer Ansätze der Mathematik in Frankreich: Poincaré, Gonthier, Cavailles
- I. HRONZKY, Changing Epistemological Perspectives in Explaining Scientific Cognition
- K. IERODIAKONOU, Preliminary Considerations on the Relation Between Stoic Logic and Greek Mathematics
- M. IGOV, Philosophical and Methodological Problems of Physical Sciences in Bulgarian Philosophical Literature After 1944
- Yu.P. IVANOV, The Formulation of Logical-Methodological Problems in Meinong's Theory of Objects
- J.K. KADRAKUNOV, Critique of Cartesian Linguistics' Methodological Foundations
- M.-L. KAKKURI-KNUUTTILA, A New Reconstruction of the Concept of Science in Aristotle
- I.T. KASAVIN, G. Berkeley as a Critic of Science and Methodology
- V.N. KATASONOV, Context of Theory
- A. KHANNA, An Alternate Perspective on the Theory of Body
- M.A. KISSEL, Dialectics of Reductionism and Anti-Reductionism in the Development of Scientific Knowledge
- W. KLEVER, Modern Aspects of Spinoza's Axiomatic Method
- O. KOISTINEN, Spinoza's Conception of Necessity and Possibility in the Ethics
- L.M. KOSAREVA, On the Value Ladenness of the Scientific Knowledge
- M.S. KOZLOVA, Methods of Language Analysis in the Late Wittgenstein's Concept
- S.B. KRIMSKY, Cultural and Historic Determinants of Science
- N.I. KUZNETSOVA, Methodology of "Presentism" in the History of Science
- J-P. LARTHOMAS, Fonction Heuristique de la loi de Continuité, d'Après la Doctrine Kantienne des Principes
- E. LASHCHYK, And not Arthur Fine's Anti-philosophical Position "NOA" either
- S.D. LATUSHKIN, N.L. MUSKHELISHVILI, On a Method of Analysing the Formation Process of Scientific Theories
- W. LENZEN, Leibniz on How to Derive Set-Theory from Elementary Arithmetics
- C.A. LERTORA MENDOZA, Robert Grosseteste: Operations and Infinite Sets
- LI HONG, CHEN TI DIAN, Four Worlds Rather Than Three—on Popper's Thesis Differentiating the Three Worlds
- LIANG QINGQIANG, The Contradictory Monism of Verification and Falsification
- J. LOSEE, The Descriptist Turn in Philosophy of Science
- S.A. LOSHCHAKOVA, On the Pre-History of the Formation of Marxism
- A. LUGG, Dehemian Realism
- L.A. MARKOVA, Empiricism and the Case Studies
- A. MATE, *Gedanke* as Statement and *Gedanke* as Proposition
- S. MCCALL, Laws of Nature and Nomic Necessity
- E. MCMULLIN, Did Newton Explain Motion?
- E. MESIMAA, An Attempt to Criticize the Concept of "Simplicity" of Classical Science
- T. MESSENGER, Lewis Carroll, John Venn, and Topological Models
- Z.N. MIKELADZE, Aristotle's Three Principles of the Methodology of Deductive Sciences
- M.I. MIKESHIN, The Problem of Methodological Analysis of Newton's Concept. Natural Philosophy Background
- L.A. MIKESHINA, Implicit Components in the Structure of Cognition
- N. MILCOV, The Unity of Wittgenstein's Philosophy of Science (Attempt for Reconstruction)
- V. MISHEVA, On the Matter of Methodology Structure

- R. MOCEK, Biologiegeschichtliche Argumentationen zur Friedensidee
 N.J. MOUTAFAKIS, Understanding the Complexities of Prohairesis Logics
 Ju.A. MURAVIEV, Problem of Truth and "The Philosophy of Science" History
 V.I. OMELIANCHIK, On Modal Paradigm of the 13th Century
 S. ONODY, On Zeno's Topos Argument
 C.W. PAGE, The Interconnections of Philosophy, the Social, Natural and Engineering Sciences and Social Practice
 B. PAHI, Independence Proofs in Propositional Logics Unaccomplishable by Finite Matrices and Decidability Without Finite Model Property
 E. PANOVA, Kant's "Transcendental Logic" as a First Attempt at a Philosophy of Science
 B.A. PARAKHONSKY, On Prehistory of Methodological Knowledge: Semiotic Models in Research Activity
 R.M. PLECKAITIS, The Theory of Propositional Equipollence in Medieval Logic
 H. POLDRACK, The Emergence of the Science of Science and the Historio-Sociological Turn in the Non-Marxist Philosophy of Science — an Example of Revolution in Social Science?
 V.N. PORUS, Methodology of Science as a Theoretical Self-Consciousness of Science
 M. PRZELECKI, Mereological Interpretation of some Paradoxes in Plato's "Parmenides"
 V.L. RABINOVICH, Towards the Problem of Historical Reconstruction of Pierre Abelard's Textology
 L.A. RADZIOVSKY, The Initial Paradox and Research Strategies in History of Psychology
 A.K. RAI, Some Remarks on the First Natural Number in Indian Logic and Mathematics
 M. REMMEL, The Rise of Developmental Biology and the 19th Century Scientific Revolution in the Biology
 V.B. RODOS, G.S. BARANOV, V.A. KOLPAKOV, The Logical Means of Theoretization of History
 V.N. SADOVSKY, Philosophy of Science in the 20th Century: Struggle of Formalistic and Antiformalistic Conceptions
 S.B. SAVENKO, Correlation of Concepts of Possibility and Reality in Aristotelian Theory of Modalities
 M.M. SHULMAN, The Solution of a Scientific Problem as Form of Transformation of Methodological Regulatives (The Example of Sadi Carnot's Heritage)
 V.S. SHYROKOV, The Medieval Mathematics and Its Influence on Galileo, Nicholas of Cusa, Leibniz and G. Cantor
 S. SLAVCOV, The Operative Role of some Symbols in the Mathematics. Marx's Conception of the Differential as an Operative Symbol
 V.A. SMIRNOV, Logical Ideas of N.A. Vasiliev and Modern Logic (invited speaker)
 D. SPASSOV, Reversed Reductionism
 B.A. STAROSTIN, J.G. Herder's Philosophy of Science and Its Connection with the Enlightenment's Conception of History
 B.V. SUBBARAYAPPA, Conception of Substance of the Indian Nyaya-Vaisesika System
 K. SUNDARAM, Narration and Explanation in History of Science
 C. THIEL, Scrutinizing an Alleged Dichotomy in the History of Mathematical Logic
 V. TISHCHENKO, Darwinism and Biological Theory
 D. TSATSOV, On the History of the Constructive Trend in Mathematics
 V.P. VIZGIN, Qualitativisme Aristotélien: Genèse et Structure
 M. V. VOLKENSHEIN, Aesthetics of Science
 S.N. VOVK, Towards the Problem of the Growing Unification in Methodology and Technology of Scientific Research
 C.A. WESTERLUND, Analysis of D'Espagnat's "Veiled Reality"
 D. WITTICH, Ernst Mach als Gesellschaftstheoretiker

- G. WOLTERS, Phenomenalism, Relativity and Atoms. Rehabilitating Ernst Mach's Philosophy of Science (invited speaker)
- V.A. YAKOVLEV, De la Détermination Sociale du Style de la Créativité Scientifique
- D. YANEVA, Science as a Self-Reproducing System. Outlines of Theoretic Description
- V.S. YAROSHENKO, A Retrospective View on Mechanicism in Relation to Evolutionism
- YIN DENG-XIANG, Some Philosophical Problems of Modern Cosmology
- YIN YONGSHENG, LIU YUESHENG, Scientific Cognition and Information Paradigm
- B.G. YUDIN, Demarcation Problem in 20th Century Methodology
- N.S. YULINA, Physicalist and Biologist Types of Metaphysics: Differences and Similarities
- E. ZARNECKA-BIALY, Why not the Fourth Figure: Some Historical Remarks on the Syllogistic Semantics
- N.Ts. ZHAMBALDAGBAJEV, The Methods of Representing Knowledge in the Tibetan Medicinal Canon "Rgyud-Bzhi"
- ZHANG JIALONG, On Aristotle's Categorical Syllogistic
- ZHANG JIALONG, Russell's Theory of Induction
- YINGBO ZHAO, New Ideas and Force Guiding the Development of Science and Technology

* * *

- V.Ya. BARKALOV, V.G. Belinskii o filosofskikh osnovaniyakh yestestvoznaniya
- A.A. KHAGUROV, Vzaimosvyaz teorii i metoda v filosofii
- S.S. KHALILOV, Istoricheskoye razvitiye vzaimosvyazi nauki s praktikoi i protsess formirovaniya ierarkhicheskoi struktury znaniya
- A.A. KHAMIDOV, Kontseptualnoye oformleniye nauki i filosofii novogo vremeni i germeticheskaya traditsiya
- Yu.V. KRYANEV, Ekumenicheskaya interpretatsiya roli nauki v tselyakh razvitiya
- Ye.V. KUZINA, K voprosu ob ideino-teoreticheskikh istochnikakh postpozitivizma
- S.A. LEBEDEV, Osnovnye linii razvitiya ponyatiya "induktsiya"
- Ya.A. MATVISHIN, Rannii etap rasprostraneniya ucheniya Kopernika v Litve, Belorussii i na Ukraine
- N.G. MIKHAI, Stanovleniye metodologii neoratsionalizma
- N.S. MUDRAGEI, Metodologiya nauki i antistsiyentizm
- L.V. POLYAKOV, Filosofiya kak nauka ili nauka kak filosofiya? (Iz istorii russkogo materializma XIX veka)
- G.G. SOLOVYOVA, Negativnaya dialektika Adorno i nauchnoye poznaniye
- V.N. YUZHAKOV, Printsip sistemnosti kak konkretno-istoricheskaya forma nauchnogo myshleniya
- K.K. ZHOL, Reprezentatsiya i referentsiya kak predmety logiko-gnoseologicheskogo analiza

INDEX OF NAMES*

- Abell, G.O. 413, 420, 424
Abraham, U. 210, 212
Abrams, M.H. 602, 605
Adam, C. 551, 624
Adler, F. 642, 643, 647, 652
Agazzi, E. 60, 83, 89
Ainslie, G. 540, 548, 550
Akaike, H. 407, 408
Akerlof, G. 535, 551
Aldous, D.J. 438, 442
Ambos-Spies, K. 201
Amellal, A. 33
Anosova, A.A. 625, 639
Apt, K.R. 108, 109
Aqvist, L. 589
Archard, G.D. 606
Aristotle 64, 68, 75, 77, 94, 109, 375
Aronson, J. 11
Arrow, K. 544, 545, 551
Arruda, A.J. 260, 625, 639
Arslanov, M.M. 191, 192, 193, 194, 199, 201
Asquith, P. 370, 410
Avenarius, R. 641, 642
Axelrod, R. 539, 542, 551
Ayer, A.J. 388
- Bacon, F. 94, 109
Baer, R. 313
Baganov, V.A. 625
Baldwin, J.T. 165, 174
Balescu, R. 439, 442
Bally, C. 606
Baltimore, D. 64, 68, 70, 77
Balzer, W. 358, 359, 363, 370
Banfield, E.G. 546, 551
Barash, D. 496, 510, 514
Barcan Marcus, R. 90, 370
- Bar-Hillel, Y. 212
Barrow, J. 421, 424
Barry, B. 534, 544, 551
Barwise, J. 105, 109, 201
Baudish 164
Baumgartner, J.E. 210, 212
Baur, W. 168, 174
Bayes 399
Bazhanov, V.A. 639
Becker, G. 534, 551
Belaval, Y. 613, 616, 624
Belegradek, O.V. 163, 174
Bellemans, A. 36, 46
Benacerraf, P. 207, 212
Bentham, J. 394
Berkeley, G. 643
Bernardo, J.M. 409, 410
Bernoulli, D. 413, 417, 424
Beth, E. 259, 260
Bickers, A. 606
Billingsley, P. 432, 442
Bismarck, O. 659
Black, M. 397, 408
Blackman, R.B. 420, 424
Blackmore, J.T. 655, 659, 660
Blair, H. 109
Blanchard, R. 382, 386
Blauberg, I.V. 340, 349
Bleek, W. 606
Block, N. 73, 74, 77
Bloomfield 600, 602
Bocheński, J.M. 640
Bogdanov, A.A. 642, 643
Bohr, N. 23, 24, 362
Bok, S. 74, 77
Boltzmann, L. 35, 36, 421, 424, 429, 442
Bondi, H. 418, 424

*The index does not cover the list of contributed papers.

- Bonik, K. 488, 492
 Bonner, J.T. 606
 Boole, G. 417, 424
 Borel, A. 173, 174
 Boring, E.G. 602, 605
 Boughattas, S. 144, 161
 Bourdieu, P. 534, 551
 Bowen, K. 301
 Bowman, P. 462, 477
 Boyd, R. 255, 260
 Boyle, R. 609
 Brentano, F. 602
 Breshnew, L.I. 658
 Bridgman 460, 462, 477
 Brout, R. 30, 41, 42, 43, 45, 46
 Browning, F. 382, 386
 Brox, O. 537
 Bruckman, G. 335, 349
 Brunshvicg, L. 612, 614, 624
 Buchholtz, W. 105, 107, 108, 109
 Buechler, S. 163, 164, 165, 168, 173,
 174
 Buhem, P. 645
 Bundy, M. 333
 Burks, A. 261
 Burrichter, C. 660
 Butler, J. 503, 514
 Butts, R.E. 644, 659
 Bühler, K. 602, 605
- Caldirola, P. 443
 Cancian, F. 536, 551
 Cantor, G. 205, 206, 207, 208, 209, 212
 Card, W. 398, 408, 410
 Carlson, L. 573, 589
 Carnap, R. 94, 109, 373, 374, 375, 383,
 387, 394, 403, 408, 443
 Carter, B. 421, 423, 424
 Carus, P. 651
 Caws, P. 375, 387
 Chadwick, L.E. 606
 Chain, E. 82, 83, 90
 Chametzky, R. 589
 Chandrasekhar, S. 36, 37, 46
 Chang, C.C. 100, 101, 109
 Chellas, B.F. 553, 567
 Cherlin, G. 165, 166, 167, 168, 174
 Chisholm, R. 384, 387
- Chomsky, N. 527, 588, 589, 591, 592,
 593, 594, 595, 597, 600, 605, 606
 Christensen, R. 399, 408
 Chuaqui, R. 247, 254, 260, 261, 639
 Church, A. 625
 Churchland, P.M. 261, 355, 356, 370
 Chwistek, L. 625
 Clark, K.L. 95, 109, 110, 282, 283, 284,
 292, 293, 297, 300, 301
 Clark 610
 Clausius, R. 429
 Codere, H. 535
 Cohen, I.B. 647, 648, 659
 Cohen, P.J. 205, 206, 207, 208, 212
 Cohen, R. 387, 388
 Cohn, P.M. 121, 138
 Collins, C.B. 422, 424
 Colodny, R.G. 261
 Colombo, P. 381, 382, 389
 Comey, D.D. 625, 640
 Commandio 610
 Cook, K. 551
 Cornfield, I.P. 432, 442
 Costantini, D. 428, 442
 Couturat, L. 615, 624
 Cresswell, M.J. 253, 260
 Crook, J.F. 395, 410
 Czermak, H. 371
- da Costa, N.C.A. 247, 249, 252, 253,
 258, 259, 260, 261, 625, 639
 dalla Chiara, M.L. 442
 Darwin, C. 63, 68, 72, 77, 482, 484,
 488, 495, 497, 509, 514, 603, 606, 638
 Dauben, J.W. 206, 208, 212
 Davidson, D. 373, 387, 388
 Davies, P.C.W. 421, 424
 Dawkins, R. 496, 500, 514
 de Finetti, B. 394, 408, 431, 438, 439,
 442
 de Lacy, E.A. 110
 de Lacy, Ph.H. 110
 de Mauro, T. 591, 606
 de Saussure, F. 591, 606
 de Vaucouleurs, G. 419, 425
 Deaton, M.L. 402, 410
 DeBuchananne, G.D. 382, 387
 Dedekind 206

- DeGroot, M.H. 409, 410
 Descartes 45, 46, 72, 327, 542, 547,
 551, 611, 612, 613, 614, 615, 624
 Deutsch, M. 538, 551
 Dewey, J. 247, 373
 Dimitracopoulos, C. 143, 152, 155, 161
 Dingle, H. 419, 424
 Dingler, H. 652, 653
 Dobzhansky 488
 Donder, H-D. 213, 214, 222
 Dorling, J. 252, 253, 260
 Dostoyevsky, F. 85
 Douglas, M. 88
 Doyle, J. 94, 110
 Drake, F.R. 201
 Dubikajtis, L. 253, 260
 Duchesneau, F. 624
 Dummet, M. 127, 138
 Durkheim, E. 531, 536
 Dworkin, G. 73, 74, 77

 Earman, J. 370
 Ebbinghaus, H.D. 201
 Eccles, J.G. 518, 527
 Eddington, A. 39, 46
 Eddy, W.F. 410
 Edgerton, R.B. 534, 535, 536, 544, 551
 Eells, E. 503, 514
 Ehlers, J. 363, 364, 370, 417, 424, 475,
 477
 Einstein, A. 24, 29, 30, 39, 41, 46, 65,
 415, 424, 429, 459, 462, 464, 647, 648,
 649, 654, 655, 659, 660
 Elcock, E.W. 409
 Eldredge, N. 489, 492
 Elitot, C.W. 109
 Ellis, B.D. 462, 477
 Elster, J. 532, 534, 539, 548, 550, 551
 Engels, F. 4, 10, 658
 Englert, F. 42, 43, 45, 46
 Epstein, R.L. 194, 201
 Ershov, Y.L. 174, 175, 194, 201
 Essler, W.K. 94, 109
 Etherington, D. 108, 109
 Evans, D. 167, 171, 174
 Everitt, C.W.F. 429, 442

 Farmer, D. 410
 Farquhar, I.E. 432, 442

 Feferman, S. 109
 Feinberg, G. 475, 477, 503, 514
 Feller, W. 435, 442
 Fenstadt, J-E. xv
 Feyerabend, P.K. 357, 362, 373, 374,
 378, 379, 380, 383, 384, 385, 387, 388
 Feynman, R.P. 18
 Field, H.H. 353, 370
 Fine, A. 257, 260
 Fine, K. 123, 139
 Fine, T.L. 394, 408
 Fisher, R.A. 401, 402, 403, 408, 417,
 424
 Fitting, M.C. 284, 301
 Flammarion, E. 639
 Føllesdal, D. 68, 74, 77, 539, 551
 Fomin, S.V. 442
 Foreman, M. 209, 224, 244
 Forrester, J.W. 334, 349
 Foucault, M. 88
 Fraenkel, A. 206, 209, 212
 Frank, P. 360
 Frank, R. 535
 Franklin, A. 381, 387
 Frazier, K. 11
 Frege, G. 573
 French, S. 247, 255, 258, 259, 260, 261
 Freud, S. 550
 Freudenthal, H. 260
 Friedberg, R. 193, 196
 Friedman, H. 41, 121, 128, 139, 143,
 161
 Friedman, M. 259, 261, 459, 477
 Frolov, I.T. xv, 23, 81, 86, 90, 175, 338,
 349, 350, 658, 659

 Gaifman, H. 362, 370
 Galavotti, M.C. 427
 Galilei, G. 9, 49, 71, 76
 Gallaire, H. 109, 110, 301
 Galtung, J. 346, 349
 Gambetta, D. 531, 551
 Gärdenfors, P. 67, 78
 Gardiner, P. 165
 Gardner, A. 594, 595
 Gardner, B. 594, 595
 Gaskins, R.A. 402, 410, 411
 Gavroglu, K. 371
 Géhéniau, J. 30, 41, 42, 46

- Gelfond, M. 301
 Geller 421
 Gelovani, V.A. 340, 349
 Gels, G. 386
 Geraets, T. 387
 Gessen, S.I. 639
 Geyzer, J. 639
 Gibbs, J.W. 399, 408, 409, 429, 442
 Giere, R.N. 364, 370, 373, 384, 387, 659
 Gigerenzer, G. 442
 Gilson, E. 624
 Glezer, V.D. 527
 Godambe, V.P. 409
 Gödel, K. 205, 206, 207, 208, 212
 Gogol, N. 638
 Goguen, J.A. 101, 109
 Goldblatt, R. 316
 Goldhaber, A. 475, 477
 Goldman, A.I. 374, 384, 387, 560, 567
 Golius 612
 Good, I.J. 66, 67, 78, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 406, 407, 408, 409, 410, 411
 Gottinger, H. 552
 Gould, S. 485, 492, 510, 514
 Gouldner, A. 346, 535, 551
 Grad, H. 441, 442
 Graham, L.R. 64, 78
 Greniewski, H. 640
 Grunbaum 459
 Gruska, J. 201
 Gunzig, E. 30, 41, 42, 43, 45, 46
 Gurney, E. 414, 424, 425
 Gutmann, W.F. 488, 492
 Gutting, G. 384, 386
 Guttman, A. 551
 Gvishiani, J.M. 340, 349

 Haas, R. 201
 Hacking, I. 257, 261, 414, 421, 422, 423, 424, 425
 Hadamar 209
 Haeckel, E. 483, 606
 Haldane, E. 624
 Hale, A. 382, 387
 Hall, G. 459
 Haller, R. 659

 Hamilton, W. 549, 551
 Handler, F. 333
 Hardy, E. 382, 387
 Hare, R.M. 80, 81, 90, 379
 Harper, W. 410, 443
 Harre, R. 11, 20
 Harrington, L. 174, 223, 225, 226, 227, 228, 231, 243, 244
 Harrop, R. 125, 139
 Hart, H.L.A. 566, 567
 Hartkaemper, A. 370
 Hartshorne, C. 247, 248, 261
 Hasenöhrl, F. 442
 Hawkin, S. 422, 423, 424
 Head 524
 Heathcote, A. 459
 Hegel, F. 3, 324, 387
 Heisenberg, W. 4, 24
 Hellman, G. 355, 356, 370, 378, 380, 387
 Helmholtz, H. 466, 467, 471
 Hempel, C. 373, 374, 375, 378, 381, 383, 384, 385, 387
 Henkin, L. 574, 589
 Hennig, W. 483, 488, 492
 Henrich 387
 Herneck, F. 642, 648, 660
 Herrera, A.D. 335
 Herrmann, E. 179, 187, 189, 190
 Herschel, W. 417, 418, 425, 428
 Heyting, H. 139, 625, 627
 Hiebert, E.N. 660
 Hilbert 303
 Hilpinen, R. xv, 371
 Hintikka, J. 369, 443, 571, 573, 574, 575, 577, 578, 580, 581, 588, 589
 Hobbes, T. 617
 Hohenstein, E. 602, 606
 Hollinger, R. 387, 388
 Holmström, G. 371, 567
 Holton, G. 77, 78, 87, 90
 Holton, H. 388
 Hooke, R. 609
 Hooker, C.A. 261
 Hopf 440
 Horwich, P. 257, 258, 261
 Hrushovski, E. 163, 164, 169, 171, 172, 173, 174

- Hubble, E. 418, 419, 422, 425
 Hubel 518
 Huchra, J. 415, 425
 Hughes, C.H. 253, 261
 Hull, D.L. 488, 489, 493, 603, 606
 Hume, D. 94, 109
 Husserl, E. 602
 Huygens, C. 610

 Ianovskiy, V. 640
 Ilyin, V. 641
 Inhetveen, R. 660
 Intriligator, M.D. 551

 Jäger, G. 106, 108, 109
 Jakobson, R. 525, 602, 606
 James, W. 247, 377, 540
 Jammer, M. 640
 Jaškowski, S. 253, 261
 Jeffreys, H. 394, 396, 397, 398, 400, 403, 411
 Jensen, A.R. 73, 78
 Jockusch, C.G. Jr. 192, 193, 194, 195, 196, 197, 199, 201
 Johnson, R. 114, 120
 Joliot-Curie, F. 24
 Jones, A.J.I. 371, 563, 564, 567, 571, 589
 Jonsson, B. 124, 139

 Kahn, A. 538, 551
 Kanger, H. 559, 565, 567
 Kanger, S. 556, 559, 562, 565, 566, 567
 Kant, I. 45, 46, 49, 76, 78, 387, 646
 Kaplan, D. 301, 573
 Kautsky, K. 642
 Kechris, A. 224, 225, 226, 244
 Kedrov, B.M. 6
 Keisler, H.J. 100, 101, 109
 Kellog, V.L. 489, 493
 Kemble, E.C. 394, 411
 Kempgen, S. 659
 Kestemont, E. 32, 37, 46
 Keynes, J.M. 109, 394, 395, 411
 Khintchine 440
 Kirby, L. 143, 161
 Kitcher, P. 410, 549, 551
 Klein, M.J. 660

 Klemke, E. 387, 388
 Kline, A. 387, 388
 Kline, G. 625, 640
 Knill-Jones, R.P. 398, 411
 Knoll, K. 382, 387
 Koerner, K. 591, 606
 Kolčinskij, E.I. 490, 491, 493
 Kolde, H. 386
 Komjáth, P. 214, 218
 Kondakov, N.I. 640
 Kopnin, P.V. 625, 640
 Koptsik, V.A. 604, 606
 Korcik, A. 640
 Korsakov, D.A. 639
 Kötter, R. 660
 Kowalski, R. 301
 Krajewski, W. 363, 364, 370
 Kramer, R. 201
 Krey, P. 382, 387
 Kripke, S. 121, 122, 123, 139, 353, 468
 Krishnaiiah, P.R. 409
 Krüger, L. 442
 Kučera, A. 194, 195, 197, 201
 Kuhn, O. 481, 493
 Kuhn, T.S. 351, 352, 354, 356, 357, 359, 360, 361, 362, 364, 368, 370, 371, 384, 388
 Kulas, J. 571, 575, 589
 Kulka, T. 378, 387
 Kunen, K. 284, 301
 Kurchatov, I.V. 24
 Kurtz, S. 197
 Kuznetsov, A.V. 121, 128, 136, 137
 Kuznetsov, B. 29, 30, 45, 46, 648, 660
 Kuznietsov, I.V. 362
 Kyburg, H.E. 408, 411, 442

 Lachlan, A.H. 165, 166, 174, 179, 180, 181, 182, 187, 190
 Ladosz, J. 640
 Lamm, H. 551
 Landmark, K. 418, 425
 Lapedes, A. 410
 Lapin, N.I. 337, 340, 349
 Laplace, P.S. 403
 Lapshin, I.N. 640
 Laudan, L. 355, 369, 370, 373, 384, 387
 Laurin, U. 544, 551

- Leibin, V.M. 346, 349
 Leibniz, G.W. 327, 465, 466, 609, 610,
 613, 614, 615, 616, 618, 620, 621, 622,
 623, 624
 Leinfellner, W. 552
 Lemaitre 41
 Lemmon, E.J. 122, 124, 127, 138, 139
 Lenin, V.I. 5, 10, 17, 492, 493, 641,
 642, 649, 658, 660
 Leonard, T. 396, 402, 410, 411
 Leontief, W. 335
 LePore, E. 388, 589
 Lerman, M. 179, 182, 190, 201
 Lessan, H. 152, 161
 Levi, I. 427
 Levi, M. 551
 Levinski, J-P. 213, 214, 222
 Levy, A. 212
 Levy, R. 541, 551
 Lewis 139
 Lewis, R.M. 441, 442
 Lewontin, R. 510, 514
 Lifschitz, V. 108, 109, 110, 301
 Lighthill, J. 34, 46
 Lindahl, L. 559, 565, 566, 567
 Lindley, D.V. 399, 409, 410, 411
 Lloyd, J.W. 95, 110, 264, 301
 Lobachevsky, N.J. 632
 Locke, J. 603, 609, 610
 Lodge, O. 414
 Longair, M.S. 424
 Lorentz, H.A. 429
 Łós, J. 124, 125, 139
 Lossky, N.O. 640
 Löther, R. 481, 493
 Lovejoy, A.D. 603, 606
 Lucasiewicz 625, 627
 Lucretius 29, 30, 46
 Ludwig, G. 363, 370, 451, 457
 Lulle, R. 614
 Luntley, M. 257, 261
 Luria, A.R. 524, 527
 Luzin, N.A. 625

 Maass, W. 179, 181, 190
 MacDowell, R. 143, 161
 Mach, E. 459, 470, 471, 641-660
 Mach, L. 649, 650, 651, 652, 653, 654

 Machiavelli 49
 MacIntyre, A. 174
 Magidor, M. 224, 225, 244
 Maher, J.P. 606
 Mahr, B. 98, 101, 110
 Makowsky, J.A. 98, 101, 110
 Maksimova, L.L. 127, 139
 Malament, D. 460, 477
 Maltsev, A.I. 98, 101, 110, 625, 640
 Marc-Wogan, K. 567
 Mardirossian 420
 Mardsen, D. 533, 551
 Mareschal, M. 32, 33, 37, 46
 Margenbesser, S. 261
 Marker, D. 225, 244
 Martin 181
 Martin, D.A. 199, 201, 224, 225, 226,
 240, 244
 Martin, M. 387
 Marx, K. 4, 10, 395, 484, 658
 Mason, J. 46
 Maxwell, J.C. 429, 442
 May, E. 492, 493
 Mayer, R. 644
 Maynard Smith, J. 500, 514
 Maynard, P. 371
 Mayr, D. 363, 370
 Mazur, P. 46
 McCarthy, J. 96, 102, 108, 110
 McClendon, H.R. 382, 387
 McDermott, D. 94, 110
 McGuinness, B. 660
 McKim, V. 384, 386
 McKinsey, J. 122, 139
 McLaughlin, R. 477
 McMullin, E. 373, 374, 375, 383, 385,
 387
 Meadows, D. 334, 335, 349
 Mendel, G.J. 482
 Menger, A. 647
 Mercer, E. 388
 Mercer, R. 109
 Messick, D.M. 538, 551
 Meyer, G. 381, 382, 388
 Meyer, S. 655, 656, 660
 Michell, J. 413, 417, 418, 419, 421, 425
 Michie, D. 409
 Mikenberg I. 247, 251, 261

- Mikula, G. 538
Minker, J. 108, 109, 110, 301
Minkowski, H. 649
Minsky, M. 523, 524, 525, 527
Mints, G.E. 121, 139
Mittelstrass, J. 643, 660
Moghissi, A. 382, 388
Montgomery, D. 386
Moore, G.H. 209, 212
Moore, R.C. 94, 110, 301
Morgan, M. 443
Morgenstern, O. 571, 589
Morison, R.S. 66, 69, 77, 78, 87, 90
Morozov, V.V. 640
Mortensen, C. 459
Moschovakis, Y. 105, 106, 110, 224, 225, 243, 244
Moulines, C-U. 363, 370
Mountcastle, V.B. 518, 527
Müller, G.H. 201
Myers, F.H. 424, 425
- Nadirov, R.F. 199, 201
Naedele, W.F. 381, 388
Nagel, E. 375, 377, 384, 388
Naqvi, S.A. 108, 110
Nardone, P. 30, 41, 42, 43, 45, 46
Narkiewicz, W. 172, 174
Nelson, R. 551
Nerlich, G. 460, 464, 469, 475, 477
Neubold, P. 382, 388
Newell, R. 380, 388
Newman, J.R. 411
Newton, I. 34, 362, 465, 466, 468, 470, 609, 610
Neyman, J. 399, 401, 403, 419, 425
Nickles, T. 256, 261, 370
Nieto, M. 475, 477
Niiniluoto, I. 351, 369, 370, 371, 589
North, D. 545, 551
Nugayev, R. 260, 261
- Ogden, C.K. 394, 411
Olech, P. 174
Oppenheim, F.E. 562, 567
Orban, J. 36, 46
Ornstein, D.S. 432, 442
Orthner, R. 656
- Ostriker, J.P. 416, 425
Owen, G. 571, 589
Ozbekhan, H. 334
- Packard, N. 410
Pappus 610, 612
Parikh, R. 94, 110
Paris, J. 112, 114, 116, 120, 143, 152, 155, 161
Parker, S. 301
Parsons, T. 536
Partridge, R.B. 419, 425
Passmore, R. 408
Pasteur, L. 482
Paulhan, F. 639
Pauli, T. 589
Pauling, L. 24
Peacock, J.A. 416, 425
Pearce, D. 356, 358, 364, 366, 368, 370, 371
Pearce, G. 371
Pearson, E.S. 399, 401, 402, 411
Peccei, A. 333, 334, 350
Peebles, P.J.E. 425, 420, 421, 425, 426
Peirce, C.S. 247, 248, 261, 398, 411, 414, 420, 425
Penzias, A.A. 419, 425
Perks, W. 396
Perrin, J-B. 655
Pettersson, I. 71, 78
Petrosky, T. 30, 35, 39, 41, 46
Petrov, M. 84, 90
Petzoldt, J. 650, 652, 653
Pfeiffer, H. xv
Philodemus 94, 110
Piddocke, S. 535, 551
Pietschmann, H. 8
Pillay, A. 163, 164, 174
Pinder, J.E. 382, 388
Pirani, F.A.E. 475, 477
Pitt, J.C. 409
Planck, M. 647, 660
Plato 592
Plechanoff, G.V. 642
Podmore, F. 424
Pohlars, W. 109
Poincaré, H. 35, 36, 208, 429, 442, 639
Poisson, S.D. 394, 398, 410, 411

- Poizat, B. 173, 174
 Pollen 520
 Popov, V.M. 637
 Popper, K.R. 41, 45, 46, 94, 110, 373, 377, 380, 388, 406, 411, 492, 493
 Pörn, I. 556, 559, 562, 564, 567
 Post, E. 625, 627
 Post, H. 364, 371
 Povarnin, S.I. 640
 Powell, C. 24
 Präger, P. 493
 Premack, D. 594, 595, 606
 Price, K.R. 382, 388
 Prigogine, I. 6, 19, 30, 31, 35, 41, 46
 Przełęcki, M. 357, 371
 Putnam, H. 207, 212, 255, 261, 353, 371, 373, 388, 575, 589
- Quattrone, G. 546, 551
 Quetelet, A. 428
 Quine, W. 360, 378, 381, 388, 460
 Quinn, P. 384, 385, 386
- Rae, A. 37, 38, 46
 Ramsey, F.P. 394, 411
 Rantala, V. 352, 364, 365, 366, 368, 371
 Raspail, F.V. 482
 Redhead, M. 259, 260, 261
 Redi, F. 482, 487
 Refsdal, S. 416, 425
 Reichenbach, H. 459, 462
 Reiter, R. 94, 95, 96, 97, 98, 100, 101, 108, 109, 110
 Reiter, R. 279, 283, 292, 293, 301
 Rhodes, D. 382, 388
 Rice, O. 535, 551
 Richardson, J. 335, 349
 Richert, H. 492
 Riedl, R. 488, 493
 Rieppel, O. 481, 492, 493
 Robb, A.A. 305, 460, 462
 Roberts, B. 335
 Robson, J.S. 408, 421, 425
 Rogers, H. 183, 190
 Rolf, B. 77
 Rolfes, E. 109
 Romney, E. 382, 388
 Rönner 520
- Rorty, R. 373, 388
 Rosen, N. 473, 477
 Rosner, F. 545, 551
 Ross, G.R.T. 624
 Ross, W.D. 77
 Ross, I. 624
 Rousseau 85
 Rován, B. 201
 Rubin, M. 212
 Rudner, R. 375, 379, 388
 Ruge, A. 638
 Russell, B. 573, 626
 Rybakov, V.V. 127, 128, 130, 131, 132, 133, 136, 137, 139
- Saarinen, E. 571, 585, 589
 Sacks, G.E. 201
 Sadovsky, V.N. 340, 349
 Sahlin, N.E. 67, 77, 78
 Salmon, W. 255, 261, 404
 Sami, R. 224, 243, 244
 Sandu, G. 560, 567, 578, 589
 Sartre, J-P. 551
 Savage, L.J. 394, 411
 Schaefer, F. 358, 371
 Schaffner, K. 388
 Scheffler, I. 373, 380, 381, 388
 Scheibe, E. 363, 371
 Schelbert, C. 606
 Schelbert, T. 606
 Schelling, T.S. 540
 Schild, A. 475, 477
 Schleicher, A. 591, 592, 606
 Schlick, M. 373, 388
 Schlipf, J.S. 108, 110
 Schmidt, H-J. 370
 Schneider, P. 416, 424, 425
 Schneider, R. 589
 Schroeder-Heister, P. 358, 371
 Schwarz, G. 407, 411
 Scott, D.S. 195, 201
 Scriven, M. 374, 375, 377, 378, 388
 Secheyaye, A. 606
 Segerberg, K. 127, 139
 Selby-Bigge, L. 514
 Sellars, W. 374, 389
 Selman, A. 126, 139

- Selten, R. 538, 552
Semenov, N.N. 4
Sentis, K. 538, 551
Shapere, D. 256, 261, 353, 371, 373,
374, 384, 385, 389
Shehtman, V.B. 128, 139
Shelah, A.H. 165, 174
Shelah, S. 212, 244
Shepherdson, J.C. 95, 96, 98, 109, 110,
144, 161, 264, 275, 297, 301
Shimony, A. 259, 261, 427
Shore, J. 114, 120
Shore, R.A. 190, 201
Shubnikov, A.V. 604, 606
Sieg, W. 109
Siegel, H. 373, 379, 384, 386, 389
Simon, H. 533
Simpson, A.W.B. 534, 552
Simpson, S.G. 195, 201
Sinai, Ya.G. 442
Singer, P. 511, 514
Sinsheimer, R.L. 65, 66, 72, 73, 78
Sjöberg, M. 567
Sklar, L. 467, 473, 477
Skyrms, B. 410, 443
Šmal'gauzen I.I. 490, 493
Smirnov, K.A. 640
Smirnov, V.A. 625, 639, 640
Smith, A. 49, 531
Smith, A.F.M. 409, 410
Smokler, H.E. 408, 411, 442
Smoluchowski 37
Sneath, P.H.A. 486, 487, 493
Sneed, J.D. 364, 370, 371
Soare, R.I. 179, 189, 190, 192, 193,
194, 195, 196, 197, 201
Sober, E. 501, 510, 514
Sokal, R.R. 486, 487, 493
Solojev, V.S. 639
Solov'ev, V.D. 199, 201
Solovay, R.M. 195, 201
Solow, R. 535, 552
Specker, E. 143, 161
Spiegelhalter, D.J. 398, 411
Spinoza 648
Sprott, D.A. 409
Stadler, F. 659
Steel 224
Stegmüller, W. 109, 355, 356, 358, 359,
363, 364, 371
Stengers, I. 31, 34, 46
Steprans, J. 212
Sternberg, S. 440, 443
Stigler, S.M. 403, 411
Straffa, P. 109
Styazhkin, N.I. 640
Suppe, F. 257, 259, 261
Suppes, P. 89, 259, 261, 371, 393, 411
Suszko, R. 124, 139
Swinburne, A.C. 638
Szasz, T.S. 660

Tannery, P. 551, 624
Tarski, A. 122, 124, 139, 247, 249
Taylor, M. 542, 552
ter Meulen, A. 573, 589
Terrace, H.S. 595, 606
Thaler, R. 534, 552
Thatcher, J.W. 109
Thomas Aquinas 49
Thomas, L. 70, 76, 78
Thomason, S.K. 123, 139
Thompson D'Arcy, W. 604, 606
Thompson, F. 355, 356, 370
Tibbetts, P. 379, 389
Tipler, F.J. 424
Tits, J. 173, 174
Tokin, B.P. 491, 493
Tolman, R.C. 429, 443
Tolstoy, L. 85, 639, 659, 660
Topor, R.W. 95, 110
Tranøy, K.E. 81, 90
Trivers, R. 499, 500, 514, 549, 552
Truesdell, C. 441, 443
Truss, J. 201
Tschulok, S. 481, 493
Tsitkin, A.I. 121, 126, 139
Tuite, K. 589
Tukey, J.W. 420, 424
Tuomela, R. 370, 371
Turing, A.M. 398, 411
Turnbull, C. 537, 552
Turner, E.L. 416, 425
Tversky, A. 534, 546, 551
Tymieniecka, A-T. 640

- Ullian, J.S. 378, 381, 388
 Ullman-Margalit, E. 545, 552
 Urban, B. 660

 van Fraassen, B. 255, 256, 257, 259,
 261, 364, 370
 Vasiliev, N.A. 625-640
 Vencovská, A. 112, 114, 116, 120
 Verharn, E. 625, 638
 Vernadsky, V.I. 5, 19, 344
 Veyne, P. 539, 552
 Vietri, M. 416, 425
 Virchow, R. 482
 Voltaire 596
 von Frisch, K. 594, 606
 von Mises, R. 430, 442
 von Neumann, J. 440, 442, 571, 589
 von Plato, J. 429, 437, 438, 439, 442,
 443
 von Smoluchowski, M. 429, 443
 von Weizsäcker, C.F. 261

 Wagner, R. 109
 Wald, A. 399, 411
 Walker, A. 109
 Wallace, A. 382, 387
 Waller, W. 537, 552
 Wallis 610
 Walsh, D. 415, 425
 Walzer, M. 535, 552
 Weber, M. 534
 Weedman, D.W. 416, 426
 Weingartner, P. 175, 371
 Weiss, A. 381, 382, 389
 Weiss, P.C. 261
 Wendroff, B. 410
 Westfall, R.S. 659
 Westwell-Roper 475, 477

 Weyl, H. 36, 46
 Wheeler, J.A. 18
 Whewell, W. 644
 Whitney, W.D. 591, 606
 Wiederman, J. 201
 Wiesel 518
 Wilkie, A. 152, 155, 161
 Wilkinson, D.T. 419, 425
 Wilks 401
 Wilson, R.W. 419, 425
 Windelband, W. 492, 638
 Wisdom 377, 380
 Wittgenstein, L. 373, 377, 380
 Wittich, D. 660
 Wolff, C. 596
 Wolters, G. 655, 656, 657, 660
 Woodin 224
 Wren 610
 Wright, J.B. 109
 Wrinch, D. 397, 411
 Wrong, D. 536, 552
 Wu, C.F. 410

 Yates 193
 Young 415, 426
 Yu, J.T. 420, 426
 Yudin, B.G. 23, 81, 84, 90
 Yudin, E.Y. 88, 90, 340, 349
 Yukawa 24

 Zagladin, V.V. 338, 350
 Zarnecka-Biała, E. 626
 Zawadskij, K.M. 490, 491, 493
 Zehner, H. 382, 389
 Zermelo, E. 206
 Zilber, B.I. 163, 164, 165, 167, 168,
 169, 171, 172, 173, 174, 175
 Zwicky, F. 415, 426